

Editorial Board

R. Bank

R.L. Graham

J. Stoer

R. Varga

H. Yserentant

E. Hairer
S. P. Nørsett
G. Wanner

Solving Ordinary Differential Equations I

Nonstiff Problems

Second Revised Edition
With 135 Figures

 Springer

Ernst Hairer
Gerhard Wanner
Université de Genève
Section de Mathématiques
2–4 rue du Lièvre
1211 Genève 4
Switzerland
Ernst.Hairer@math.unige.ch
Gerhard.Wanner@math.unige.ch

Syvert P. Nørsett
Norwegian University of Science
and Technology (NTNU)
Department of Mathematical Sciences
7491 Trondheim
Norway
norsett@math.ntnu.no

Corrected 3rd printing 2008

ISBN 978-3-540-56670-0

e-ISBN 978-3-540-78862-1

DOI 10.1007/978-3-540-78862-1

Springer Series in Computational Mathematics ISSN 0179-3632

Library of Congress Control Number: 93007847

Mathematics Subject Classification (2000): 65Lxx, 34A50

© 1993, 1987 Springer-Verlag Berlin Heidelberg

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: WMX Design GmbH, Heidelberg

Typesetting: by the authors

Production: LE-TeX Jelonek, Schmidt & Vöckler GbR, Leipzig

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

This edition is dedicated to
Professor John Butcher
on the occasion of his 60th birthday

His unforgettable lectures on Runge-Kutta methods, given in June 1970 at the University of Innsbruck, introduced us to this subject which, since then, we have never ceased to love and to develop with all our humble abilities.

From the Preface to the First Edition

So far as I remember, I have never seen an Author's Preface which had any purpose but one — to furnish reasons for the publication of the Book. (Mark Twain)

Gauss' dictum, "when a building is completed no one should be able to see any trace of the scaffolding," is often used by mathematicians as an excuse for neglecting the motivation behind their own work and the history of their field. Fortunately, the opposite sentiment is gaining strength, and numerous asides in this Essay show to which side go my sympathies. (B.B. Mandelbrot 1982)

This gives us a good occasion to work out most of the book until the next year. (the Authors in a letter, dated Oct. 29, 1980, to Springer-Verlag)

There are two volumes, one on non-stiff equations, . . . , the second on stiff equations, The first volume has three chapters, one on classical mathematical theory, one on Runge-Kutta and extrapolation methods, and one on multistep methods. There is an Appendix containing some Fortran codes which we have written for our numerical examples.

Each chapter is divided into sections. Numbers of formulas, theorems, tables and figures are consecutive in each section and indicate, in addition, the section number, but not the chapter number. Cross references to other chapters are rare and are stated explicitly. . . . References to the Bibliography are by "Author" plus "year" in parentheses. The Bibliography makes no attempt at being complete; we have listed mainly the papers which are discussed in the text.

Finally, we want to thank all those who have helped and encouraged us to prepare this book. The marvellous "Minisymposium" which G. Dahlquist organized in Stockholm in 1979 gave us the first impulse for writing this book. J. Steinig and Chr. Lubich have read the whole manuscript very carefully and have made extremely valuable mathematical and linguistic suggestions. We also thank J.P. Eckmann for his troff software with the help of which the whole manuscript has been printed. For preliminary versions we had used textprocessing programs written by R. Menk. Thanks also to the staff of the Geneva computing center for their help. All computer plots have been done on their beautiful HP plotter. Last but not least, we would like to acknowledge the agreeable collaboration with the planning and production group of Springer-Verlag.

October 29, 1986

The Authors

Preface to the Second Edition

The preparation of the second edition has presented a welcome opportunity to improve the first edition by rewriting many sections and by eliminating errors and misprints. In particular we have included new material on

- Hamiltonian systems (I.14) and symplectic Runge-Kutta methods (II.16);
- dense output for Runge-Kutta (II.6) and extrapolation methods (II.9);
- a new Dormand & Prince method of order 8 with dense output (II.5);
- parallel Runge-Kutta methods (II.11);
- numerical tests for first- and second order systems (II.10 and III.7).

Our sincere thanks go to many persons who have helped us with our work:

- all readers who kindly drew our attention to several errors and misprints in the first edition;
- those who read preliminary versions of the new parts of this edition for their invaluable suggestions: D.J. Higham, L. Jay, P. Kaps, Chr. Lubich, B. Moesli, A. Ostermann, D. Pfenniger, P.J. Prince, and J.M. Sanz-Serna.
- our colleague J. Steinig, who read the entire manuscript, for his numerous mathematical suggestions and corrections of English (and Latin!) grammar;
- our colleague J.P. Eckmann for his great skill in manipulating Apollo workstations, font tables, and the like;
- the staff of the Geneva computing center and of the mathematics library for their constant help;
- the planning and production group of Springer-Verlag for numerous suggestions on presentation and style.

This second edition now also benefits, as did Volume II, from the marvels of \TeX nology. All figures have been recomputed and printed, together with the text, in Postscript. Nearly all computations and text processings were done on the Apollo DN4000 workstation of the Mathematics Department of the University of Geneva; for some long-time and high-precision runs we used a VAX 8700 computer and a Sun IPX workstation.

November 29, 1992

The Authors

Contents

Chapter I. Classical Mathematical Theory

I.1	Terminology	2
I.2	The Oldest Differential Equations	4
	Newton	4
	Leibniz and the Bernoulli Brothers	6
	Variational Calculus	7
	Clairaut	9
	Exercises	10
I.3	Elementary Integration Methods	12
	First Order Equations	12
	Second Order Equations	13
	Exercises	14
I.4	Linear Differential Equations	16
	Equations with Constant Coefficients	16
	Variation of Constants	18
	Exercises	19
I.5	Equations with Weak Singularities	20
	Linear Equations	20
	Nonlinear Equations	23
	Exercises	24
I.6	Systems of Equations	26
	The Vibrating String and Propagation of Sound	26
	Fourier	29
	Lagrangian Mechanics	30
	Hamiltonian Mechanics	32
	Exercises	34
I.7	A General Existence Theorem	35
	Convergence of Euler's Method	35
	Existence Theorem of Peano	41
	Exercises	43
I.8	Existence Theory using Iteration Methods and Taylor Series	44
	Picard-Lindelöf Iteration	45
	Taylor Series	46
	Recursive Computation of Taylor Coefficients	47
	Exercises	49

I.9	Existence Theory for Systems of Equations	51
	Vector Notation	52
	Subordinate Matrix Norms	53
	Exercises	55
I.10	Differential Inequalities	56
	Introduction	56
	The Fundamental Theorems	57
	Estimates Using One-Sided Lipschitz Conditions	60
	Exercises	62
I.11	Systems of Linear Differential Equations	64
	Resolvent and Wronskian	65
	Inhomogeneous Linear Equations	66
	The Abel-Liouville-Jacobi-Ostrogradskii Identity	66
	Exercises	67
I.12	Systems with Constant Coefficients	69
	Linearization	69
	Diagonalization	69
	The Schur Decomposition	70
	Numerical Computations	72
	The Jordan Canonical Form	73
	Geometric Representation	77
	Exercises	78
I.13	Stability	80
	Introduction	80
	The Routh-Hurwitz Criterion	81
	Computational Considerations	85
	Liapunov Functions	86
	Stability of Nonlinear Systems	87
	Stability of Non-Autonomous Systems	88
	Exercises	89
I.14	Derivatives with Respect to Parameters and Initial Values ...	92
	The Derivative with Respect to a Parameter	93
	Derivatives with Respect to Initial Values	95
	The Nonlinear Variation-of-Constants Formula	96
	Flows and Volume-Preserving Flows	97
	Canonical Equations and Symplectic Mappings	100
	Exercises	104
I.15	Boundary Value and Eigenvalue Problems	105
	Boundary Value Problems	105
	Sturm-Liouville Eigenvalue Problems	107
	Exercises	110
I.16	Periodic Solutions, Limit Cycles, Strange Attractors	111
	Van der Pol's Equation	111
	Chemical Reactions	115
	Limit Cycles in Higher Dimensions, Hopf Bifurcation	117
	Strange Attractors	120
	The Ups and Downs of the Lorenz Model	123
	Feigenbaum Cascades	124
	Exercises	126

Chapter II. Runge-Kutta and Extrapolation Methods

II.1	The First Runge-Kutta Methods	132
	General Formulation of Runge-Kutta Methods	134
	Discussion of Methods of Order 4	135
	“Optimal” Formulas	139
	Numerical Example	140
	Exercises	141
II.2	Order Conditions for Runge-Kutta Methods	143
	The Derivatives of the True Solution	145
	Conditions for Order 3	145
	Trees and Elementary Differentials	145
	The Taylor Expansion of the True Solution	148
	Faà di Bruno’s Formula	149
	The Derivatives of the Numerical Solution	151
	The Order Conditions	153
	Exercises	154
II.3	Error Estimation and Convergence for RK Methods	156
	Rigorous Error Bounds	156
	The Principal Error Term	158
	Estimation of the Global Error	159
	Exercises	163
II.4	Practical Error Estimation and Step Size Selection	164
	Richardson Extrapolation	164
	Embedded Runge-Kutta Formulas	165
	Automatic Step Size Control	167
	Starting Step Size	169
	Numerical Experiments	170
	Exercises	172
II.5	Explicit Runge-Kutta Methods of Higher Order	173
	The Butcher Barriers	173
	6-Stage, 5th Order Processes	175
	Embedded Formulas of Order 5	176
	Higher Order Processes	179
	Embedded Formulas of High Order	180
	An 8th Order Embedded Method	181
	Exercises	185
II.6	Dense Output, Discontinuities, Derivatives	188
	Dense Output	188
	Continuous Dormand & Prince Pairs	191
	Dense Output for DOP853	194
	Event Location	195
	Discontinuous Equations	196
	Numerical Computation of Derivatives with Respect to Initial Values and Parameters	200
	Exercises	202
II.7	Implicit Runge-Kutta Methods	204
	Existence of a Numerical Solution	206
	The Methods of Kuntzmann and Butcher of Order 2s	208
	IRK Methods Based on Lobatto Quadrature	210

	Collocation Methods	211
	Exercises	214
II.8	Asymptotic Expansion of the Global Error	216
	The Global Error	216
	Variable h	218
	Negative h	219
	Properties of the Adjoint Method	220
	Symmetric Methods	221
	Exercises	223
II.9	Extrapolation Methods	224
	Definition of the Method	224
	The Aitken - Neville Algorithm	226
	The Gragg or GBS Method	228
	Asymptotic Expansion for Odd Indices	231
	Existence of Explicit RK Methods of Arbitrary Order	232
	Order and Step Size Control	233
	Dense Output for the GBS Method	237
	Control of the Interpolation Error	240
	Exercises	241
II.10	Numerical Comparisons	244
	Problems	244
	Performance of the Codes	249
	A "Stretched" Error Estimator for DOP853	254
	Effect of Step-Number Sequence in ODEX	256
II.11	Parallel Methods	257
	Parallel Runge-Kutta Methods	258
	Parallel Iterated Runge-Kutta Methods	259
	Extrapolation Methods	261
	Increasing Reliability	261
	Exercises	263
II.12	Composition of B-Series	264
	Composition of Runge-Kutta Methods	264
	B-Series	266
	Order Conditions for Runge-Kutta Methods	269
	Butcher's "Effective Order"	270
	Exercises	272
II.13	Higher Derivative Methods	274
	Collocation Methods	275
	Hermite-Obreschkoff Methods	277
	Fehlberg Methods	278
	General Theory of Order Conditions	280
	Exercises	281
II.14	Numerical Methods for Second Order Differential Equations	283
	Nyström Methods	284
	The Derivatives of the Exact Solution	286
	The Derivatives of the Numerical Solution	288
	The Order Conditions	290
	On the Construction of Nyström Methods	291
	An Extrapolation Method for $y'' = f(x, y)$	294
	Problems for Numerical Comparisons	296

Performance of the Codes	298
Exercises	300
II.15 P-Series for Partitioned Differential Equations	302
Derivatives of the Exact Solution, P-Trees	303
P-Series	306
Order Conditions for Partitioned Runge-Kutta Methods	307
Further Applications of P-Series	308
Exercises	311
II.16 Symplectic Integration Methods	312
Symplectic Runge-Kutta Methods	315
An Example from Galactic Dynamics	319
Partitioned Runge-Kutta Methods	326
Symplectic Nyström Methods	330
Conservation of the Hamiltonian; Backward Analysis	333
Exercises	337
II.17 Delay Differential Equations	339
Existence	339
Constant Step Size Methods for Constant Delay	341
Variable Step Size Methods	342
Stability	343
An Example from Population Dynamics	345
Infectious Disease Modelling	347
An Example from Enzyme Kinetics	248
A Mathematical Model in Immunology	349
Integro-Differential Equations	351
Exercises	352

Chapter III. Multistep Methods and General Linear Methods

III.1 Classical Linear Multistep Formulas	356
Explicit Adams Methods	357
Implicit Adams Methods	359
Numerical Experiment	361
Explicit Nyström Methods	362
Milne–Simpson Methods	363
Methods Based on Differentiation (BDF)	364
Exercises	366
III.2 Local Error and Order Conditions	368
Local Error of a Multistep Method	368
Order of a Multistep Method	370
Error Constant	372
Irreducible Methods	374
The Peano Kernel of a Multistep Method	375
Exercises	377
III.3 Stability and the First Dahlquist Barrier	378
Stability of the BDF-Formulas	380
Highest Attainable Order of Stable Multistep Methods	383
Exercises	387

III.4 Convergence of Multistep Methods	391
Formulation as One-Step Method	393
Proof of Convergence	395
Exercises	396
III.5 Variable Step Size Multistep Methods	397
Variable Step Size Adams Methods	397
Recurrence Relations for $g_j(n)$, $\Phi_j(n)$ and $\Phi_j^*(n)$	399
Variable Step Size BDF	400
General Variable Step Size Methods and Their Orders	401
Stability	402
Convergence	407
Exercises	409
III.6 Nordsieck Methods	410
Equivalence with Multistep Methods	412
Implicit Adams Methods	417
BDF-Methods	419
Exercises	420
III.7 Implementation and Numerical Comparisons	421
Step Size and Order Selection	421
Some Available Codes	423
Numerical Comparisons	427
III.8 General Linear Methods	430
A General Integration Procedure	431
Stability and Order	436
Convergence	438
Order Conditions for General Linear Methods	441
Construction of General Linear Methods	443
Exercises	445
III.9 Asymptotic Expansion of the Global Error	448
An Instructive Example	448
Asymptotic Expansion for Strictly Stable Methods (8.4)	450
Weakly Stable Methods	454
The Adjoint Method	457
Symmetric Methods	459
Exercises	460
III.10 Multistep Methods for Second Order Differential Equations	461
Explicit Störmer Methods	462
Implicit Störmer Methods	464
Numerical Example	465
General Formulation	467
Convergence	468
Asymptotic Formula for the Global Error	471
Rounding Errors	472
Exercises	473
Appendix. Fortran Codes	475
Driver for the Code DOPRI5	475
Subroutine DOPRI5	477
Subroutine DOP853	481
Subroutine ODEX	482

Subroutine ODEX2	484
Driver for the Code RETARD	486
Subroutine RETARD	488
Bibliography	491
Symbol Index	521
Subject Index	523

Chapter I. Classical Mathematical Theory

... halte ich es immer für besser, nicht mit dem Anfang anzufangen, der immer das Schwerste ist.

(B. Riemann copied this from F. Schiller into his notebook)

This first chapter contains the classical theory of differential equations, which we judge useful and important for a profound understanding of numerical processes and phenomena. It will also be the occasion of presenting interesting examples of differential equations and their properties.

We first retrace in Sections I.2-I.6 the historical development of classical integration methods by series expansions, quadrature and elementary functions, from the beginning (Newton and Leibniz) to the era of Euler, Lagrange and Hamilton. The next part (Sections I.7-I.14) deals with theoretical properties of the solutions (existence, uniqueness, stability and differentiability with respect to initial values and parameters) and the corresponding flow (increase of volume, preservation of symplectic structure). This theory was initiated by Cauchy in 1824 and then brought to perfection mainly during the next 100 years. We close with a brief account of boundary value problems, periodic solutions, limit cycles and strange attractors (Sections I.15 and I.16).

I.1 Terminology

A *differential equation of first order* is an equation of the form

$$y' = f(x, y) \quad (1.1)$$

with a given function $f(x, y)$. A function $y(x)$ is called a *solution* of this equation if for all x ,

$$y'(x) = f(x, y(x)). \quad (1.2)$$

It was observed very early by Newton, Leibniz and Euler that the solution usually contains a free parameter, so that it is uniquely determined only when an *initial value*

$$y(x_0) = y_0 \quad (1.3)$$

is prescribed. Cauchy's existence and uniqueness proof of this fact will be discussed in Section I.7. Differential equations arise in many applications. We shall see the first examples of such equations in Section I.2, and in Section I.3 how some of them can be solved explicitly.

A *differential equation of second order* for y is of the form

$$y'' = f(x, y, y'). \quad (1.4)$$

Here, the solution usually contains *two* parameters and is only uniquely determined by *two* initial values

$$y(x_0) = y_0, \quad y'(x_0) = y'_0. \quad (1.5)$$

Equations of second order can rarely be solved explicitly (see I.3). For their numerical solution, as well as for theoretical investigations, one usually sets $y_1(x) := y(x)$, $y_2(x) := y'(x)$, so that equation (1.4) becomes

$$\begin{aligned} y'_1 &= y_2 & y_1(x_0) &= y_0 \\ y'_2 &= f(x, y_1, y_2) & y_2(x_0) &= y'_0. \end{aligned} \quad (1.4')$$

This is an example of a *first order system of differential equations*, of dimension n (see Sections I.6 and I.9),

$$\begin{aligned} y'_1 &= f_1(x, y_1, \dots, y_n) & y_1(x_0) &= y_{10} \\ &\dots & \dots & \\ y'_n &= f_n(x, y_1, \dots, y_n) & y_n(x_0) &= y_{n0}. \end{aligned} \quad (1.6)$$

Most of the theory of this book is devoted to the solution of the initial value problem for the system (1.6). At the end of the 19th century (Peano 1890) it became customary to introduce the vector notation

$$y = (y_1, \dots, y_n)^T, \quad f = (f_1, \dots, f_n)^T$$

so that (1.6) becomes $y' = f(x, y)$, which is again the same as (1.1), but now with y and f interpreted as vectors.

Another possibility for the second order equation (1.4), instead of transforming it into a system (1.4'), is to develop *methods specially adapted to second order equations* (Nyström methods). This will be done in special sections of this book (Sections II.13 and III.10). Nothing prevents us, of course, from considering (1.4) as a second order system of dimension n .

If, however, the initial conditions (1.5) are replaced by something like $y(x_0) = a$, $y(x_1) = b$, i.e., if the conditions determining the particular solution are not all specified at the same point x_0 , we speak of a *boundary value problem*. The theory of the existence of a solution and of its numerical computation is here much more complicated. We give some examples in Section I.15.

Finally, a problem of the type

$$\frac{\partial u}{\partial t} = f\left(t, u, \frac{\partial u}{\partial x}, \frac{\partial^2 u}{\partial x^2}\right) \quad (1.7)$$

for an unknown function $u(t, x)$ of two *independent variables* will be called a *partial differential equation*. We can also deal with partial differential equations of higher order, with problems in three or four independent variables, or with systems of partial differential equations. Very often, initial value problems for partial differential equations can conveniently be transformed into a system of ordinary differential equations, for example with finite difference or finite element approximations in the variable x . In this way, the equation

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2}$$

would become

$$\frac{du_i}{dt} = \frac{a^2}{\Delta x^2} (u_{i+1} - 2u_i + u_{i-1}),$$

where $u_i(t) \approx u(t, x_i)$. This procedure is called the “method of lines” or “method of discretization in space” (Berezin & Zhidkov 1965). We shall see in Section I.6 that this connection, the other way round, was historically the origin of partial differential equations (d’Alembert, Lagrange, Fourier). A similar idea is the “method of discretization in time” (Rothe 1930).

I.2 The Oldest Differential Equations

... So zum Beispiel die Aufgabe der umgekehrten Tangentenmethode, von welcher auch Descartes eingestand, dass er sie nicht in seiner Gewalt habe. (Leibniz, 27. Aug 1676)

... et on sait que les seconds Inventeurs n'ont pas de droit à l'Invention. (Newton, 29 mai 1716)

Il ne paroist point que M. Newton ait eu avant moy la caracteristique & l'algorithme infinitesimal ... (Leibniz)

And by these words he acknowledged that he had not yet found the reduction of problems to differential equations. (Newton)

Newton

Differential equations are as old as differential calculus. Newton considered them in his treatise on differential calculus (Newton 1671) and discussed their solution by series expansion. One of the first examples of a first order equation treated by Newton (see Newton (1671), Problema II, Solutio Casus II, Ex. I) was

$$y' = 1 - 3x + y + x^2 + xy. \quad (2.1)$$

For each value x and y , such an equation prescribes the derivative y' of the solutions. We thus obtain a *vector field*, which, for this particular equation, is sketched in Fig. 2.1a. (So, contrary to the belief of many people, vector fields existed long before Van Gogh). The solutions are the curves which respect these prescribed directions everywhere (Fig. 2.1b).

Newton discusses the solution of this equation by means of infinite series, whose terms he obtains recursively (“... & ils se jettent sur les series, où M. Newton m’a precedé sans difficulté; mais ...”, Leibniz). The first term

$$y = 0 + \dots$$

is the initial value for $x = 0$. Inserting this into the differential equation (2.1) he obtains

$$y' = 1 + \dots$$

which, integrated, gives

$$y = x + \dots$$

Again, from (2.1), we now have

$$y' = 1 - 3x + x + \dots = 1 - 2x + \dots$$

and by integration

$$y = x - x^2 + \dots$$

E X E M P L. I

Sit Aequatio $\frac{y}{x} = 1 - 3x + y + xx + xy$, cujus Terminos:
 $x - 3x + xx$ non affectos *Relata* Quantitate dispositos vides in la-
 teralem Seriem primo loco, & reliquos y & xy in sinistrâ Columnâ.

	$+ 1 - 3x + xx$
$+ y$	$* + x - xx + \frac{1}{3}x^3 - \frac{1}{6}x^4 + \frac{1}{30}x^5; \&c.$
$+ xy$	$* x + xx - x^3 + \frac{1}{3}x^4 - \frac{1}{6}x^5 + \frac{1}{30}x^6; \&c.$
Aggreg.	$+ 1 - 2x + xx - \frac{2}{3}x^3 + \frac{1}{6}x^4 - \frac{4}{30}x^5; \&c.$
$y =$	$+ x - xx + \frac{1}{3}x^3 - \frac{1}{6}x^4 + \frac{1}{30}x^5 - \frac{1}{45}x^6; \&c.$

Nunc:

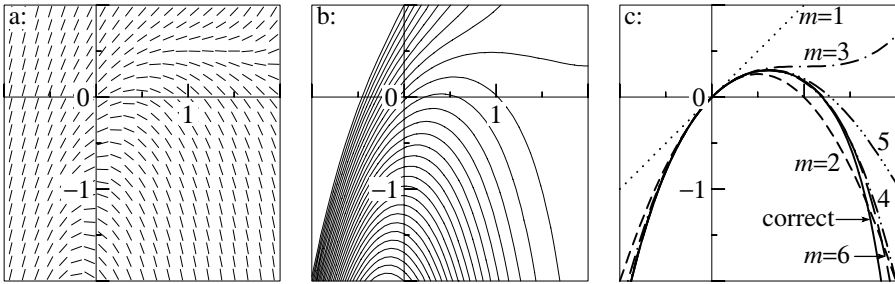


Fig. 2.1. a) vector field, b) various solution curves of equation (2.1),
 c) Correct solution vs. approximate solution

The next round gives

$$y' = 1 - 2x + x^2 + \dots, \quad y = x - x^2 + \frac{x^3}{3} + \dots$$

Continuing this process, he finally arrives at

$$y = x - xx + \frac{1}{3}x^3 - \frac{1}{6}x^4 + \frac{1}{30}x^5 - \frac{1}{45}x^6; \&c. \quad (2.2)$$

These approximations, term after term, are plotted in Fig. 2.1c together with the correct solution. It can be seen that these approximations are closer and closer to the true solution for small values of x . For more examples see Exercises 1-3. Convergence will be discussed in Section I.8.

Leibniz and the Bernoulli Brothers

A second access to differential equations is the consideration of geometrical problems such as *inverse tangent problems* (Debeaune 1638 in a letter to Descartes). A particular example describes the path of a silver pocket watch (“horologio portabili suae thecae argenteae”) and was proposed around 1674 by “Claudius Perraltus Medicus Parisinus” to Leibniz: a curve $y(x)$ is required whose tangent AB is given, say everywhere of constant length a (Fig. 2.2). This leads to

$$y' = -\frac{y}{\sqrt{a^2 - y^2}}, \tag{2.3}$$

a first order differential equation. Despite the efforts of the “plus célèbres mathématiciens de Paris et de Toulouse” (from a letter of Descartes 1645, “Toulouse” means “Fermat”) the solution of these problems had to wait until Leibniz (1684) and above all until the famous paper of Jacob Bernoulli (1690). Bernoulli’s idea applied to equation (2.3) is as follows: let the curve BM in Fig. 2.3 be such that LM is equal to $\sqrt{a^2 - y^2}/y$. Then (2.3), written as

$$dx = -\frac{\sqrt{a^2 - y^2}}{y} dy, \tag{2.3'}$$

shows that for *all* y the areas S_1 and S_2 (Fig. 2.3) are the same. Thus (“Ergo & horum integralia aequantur”) the areas $BMLB$ and $A_1A_2C_2C_1$ must be equal too. Hence (2.3’) becomes (Leibniz 1693)

$$x = \int_y^a \frac{\sqrt{a^2 - y^2}}{y} dy = -\sqrt{a^2 - y^2} - a \cdot \log \frac{a - \sqrt{a^2 - y^2}}{y}. \tag{2.3''}$$

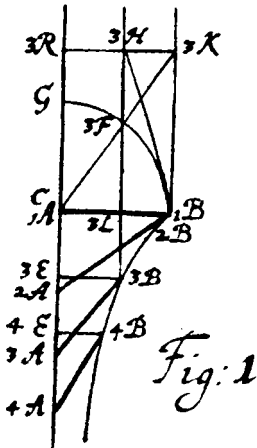


Fig. 2.2. Illustration from Leibniz (1693)

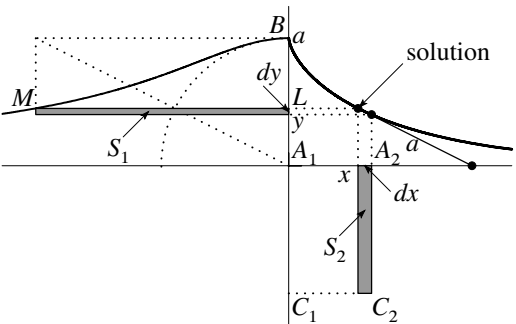


Fig. 2.3. Jac. Bernoulli’s Solution of (2.3)

Variational Calculus

In 1696 Johann Bernoulli invited the brightest mathematicians of the world (“Profundioris in primis Mathesos cultori, Salutem!”) to solve the *brachystochrone* (shortest time) problem, mainly in order to fault his brother Jacob, from whom he expected a wrong solution. The problem is to find a curve $y(x)$ connecting two points P_0, P_1 , such that a point gliding on this curve under gravitation reaches P_1 in the shortest time possible. In order to solve his problem, Joh. Bernoulli (1697b) imagined thin layers of homogeneous media and knew from optics (Fermat’s principle) that a light ray with speed v obeying the law of Snellius

$$\sin \alpha = Kv$$

passes through in the shortest time. Since the speed is known to be proportional to the square root of the fallen height, he obtains, by passing to thinner and thinner layers,

$$\sin \alpha = \frac{1}{\sqrt{1+y'^2}} = K\sqrt{2g(y-h)}, \quad (2.4)$$

a differential equation of the first order.

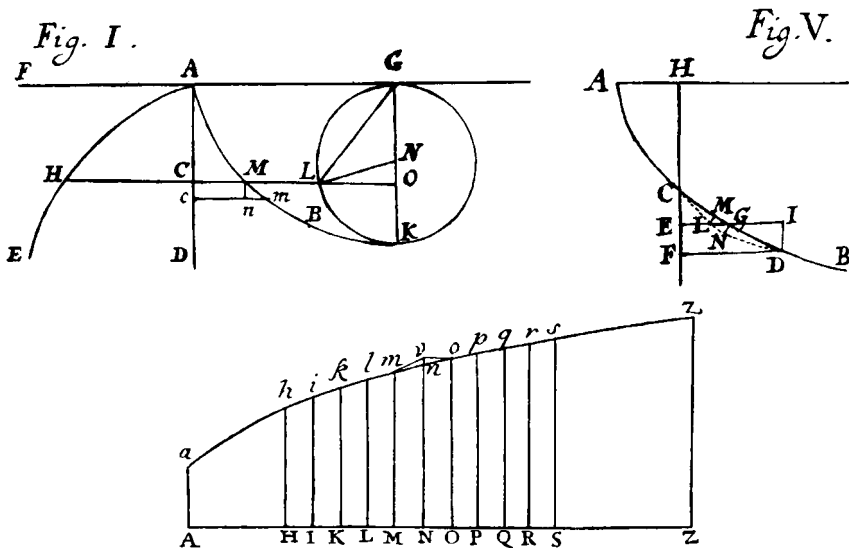


Fig. 2.4. Solutions of the variational problem (Joh. Bernoulli, Jac. Bernoulli, Euler)

The solutions of (2.4) can be shown to be cycloids (see Exercise 6 of Section I.3). Jacob, in his reply, also furnished a solution, much less elegant but unfortunately correct. Jacob’s method (see Fig. 2.4) was something like today’s (inverse)

“finite element” method and more general than Johann’s and led to the famous work of Euler (1744), which gives the general solution of the problem

$$\int_{x_0}^{x_1} F(x, y, y') dx = \min \quad (2.5)$$

with the help of the differential equation of the second order

$$F_y(x, y, y') - \frac{d}{dx} (F_{y'}(x, y, y')) = F_y - F_{y'y'}y'' - F_{y'yy'}y' - F_{y'yx} = 0, \quad (2.6)$$

and treated 100 variational problems in detail. Equation (2.6), in the special case where F does not depend on x , can be integrated to give

$$F - F_{y'}y' = K. \quad (2.6')$$

Euler’s original proof used polygons in order to establish equation (2.6). Only the ideas of Lagrange, in 1755 at the age of 19, led to the proof which is today the usual one (letter of Aug. 12, 1755; Oeuvres vol. 14, p. 138): add an arbitrary “variation” $\delta y(x)$ to $y(x)$ and linearize (2.5).

$$\begin{aligned} \int_{x_0}^{x_1} F(x, y + \delta y, y' + (\delta y)') dx \\ = \int_{x_0}^{x_1} F(x, y, y') dx + \int_{x_0}^{x_1} \left(F_y(x, y, y') \delta y + F_{y'}(x, y, y') (\delta y)' \right) dx + \dots \end{aligned} \quad (2.7)$$

The last integral in (2.7) represents the “derivative” of (2.5) with respect to δy . Therefore, if $y(x)$ is the solution of (2.5), we must have

$$\int_{x_0}^{x_1} \left(F_y(x, y, y') \delta y + F_{y'}(x, y, y') (\delta y)' \right) dx = 0 \quad (2.8)$$

or, after partial integration,

$$\int_{x_0}^{x_1} \left(F_y(x, y, y') - \frac{d}{dx} F_{y'}(x, y, y') \right) \cdot \delta y(x) dx = 0. \quad (2.8')$$

Since (2.8') must be fulfilled by all δy , Lagrange “sees” that

$$F_y(x, y, y') - \frac{d}{dx} F_{y'}(x, y, y') = 0 \quad (2.9)$$

is necessary for (2.5). Euler, in his reply (Sept. 6, 1755) urged a more precise proof of this fact (which is now called the “fundamental Lemma of variational Calculus”). For *several* unknown functions

$$\int F(x, y_1, y'_1, \dots, y_n, y'_n) dx = \min \quad (2.10)$$

the same proof leads to the equations

$$F_{y_i}(x, y_1, y'_1, \dots, y_n, y'_n) - \frac{d}{dx} F_{y'_i}(x, y_1, y'_1, \dots, y_n, y'_n) = 0 \quad (2.11)$$

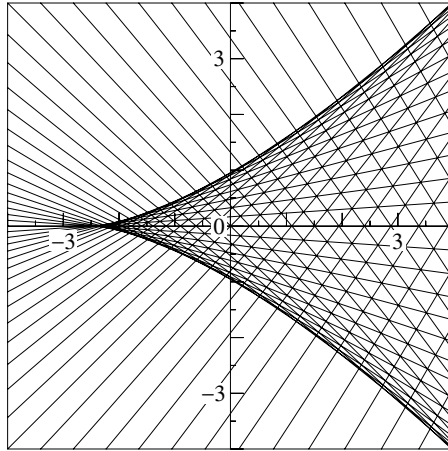


Fig. 2.6. Solutions of a Clairaut differential equation

Exercises

1. (Newton). Solve equation (2.1) with another initial value $y(0) = 1$.

Newton's result: $y = 1 + 2x + x^3 + \frac{1}{4}x^4 + \frac{1}{4}x^5$, &c.

2. (Newton 1671, "Problema II, Solutio particolare"). Solve the total differential equation

$$3x^2 - 2ax + ay - 3y^2y' + axy' = 0.$$

Solution given by Newton: $x^3 - ax^2 + axy - y^3 = 0$. Observe that he missed the arbitrary integration constant C .

3. (Newton 1671). Solve the equations

a) $y' = 1 + \frac{y}{a} + \frac{xy}{a^2} + \frac{x^2y}{a^3} + \frac{x^3y}{a^4}$, &c.

b) $y' = -3x + 3xy + y^2 - xy^2 + y^3 - xy^3 + y^4 - xy^4 + 6x^2y - 6x^2 + 8x^3y - 8x^3 + 10x^4y - 10x^4$, &c.

Results given by Newton:

a) $y = x + \frac{x^2}{2a} + \frac{x^3}{2a^2} + \frac{x^4}{2a^3} + \frac{x^5}{2a^4} + \frac{x^6}{2a^5}$, &c.

b) $y = -\frac{3}{2}x^2 - 2x^3 - \frac{25}{8}x^4 - \frac{91}{20}x^5 - \frac{111}{16}x^6 - \frac{367}{35}x^7$, &c.

4. Show that the differential equation

$$x + yy' = y'\sqrt{x^2 + y^2 - 1}$$

possesses the solutions $2ay = a^2 + 1 - x^2$ for all a . Sketch these curves and find yet another solution of the equation (from Lagrange (1774), p. 7, which was written to explain the “Clairaut phenomenon”).

5. Verify that the envelope of the solutions $y = Cx - f(C)$ of the Clairaut equation (2.12) is given in parametric representation by

$$x(p) = f'(p)$$

$$y(p) = pf'(p) - f(p).$$

Show that this envelope is also a solution of (2.12) and calculate it for $f(C) = 5(C^3 - C)/2$ (cf. Fig. 2.6).

6. (Cauchy 1824). Show that the family $y = C(x + C)^2$ satisfies the differential equation $(y')^3 = 8y^2 - 4xyy'$. Find yet another solution which is not included in this family (see Fig. 2.7).

Answer: $y = -\frac{4}{27}x^3$.

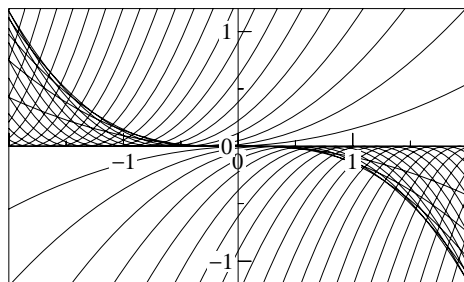


Fig. 2.7. Solution family of Cauchy's example in Exercise 6

I.3 Elementary Integration Methods

We now discuss some of the simplest types of equations, which can be solved by the computation of integrals.

First Order Equations

The equation with separable variables.

$$y' = f(x)g(y). \quad (3.1)$$

Extending the idea of Jacob Bernoulli (see (2.3')), we divide by $g(y)$, integrate and obtain the solution (Leibniz 1691, in a letter to Huygens)

$$\int \frac{dy}{g(y)} = \int f(x) dx + C.$$

A special example of this is the *linear equation* $y' = f(x)y$, which possesses the solution

$$y(x) = CR(x), \quad R(x) = \exp\left(\int f(x) dx\right).$$

The inhomogeneous linear equation.

$$y' = f(x)y + g(x). \quad (3.2)$$

Here, the substitution $y(x) = c(x)R(x)$ leads to $c'(x) = g(x)/R(x)$ (Joh. Bernoulli 1697). One thus obtains the solution

$$y(x) = R(x) \left(\int_{x_0}^x \frac{g(s)}{R(s)} ds + C \right). \quad (3.3)$$

Total differential equations. An equation of the form

$$P(x, y) + Q(x, y)y' = 0 \quad (3.4)$$

is found to be immediately solvable if

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x}. \quad (3.5)$$

One can then find by integration a potential function $U(x, y)$ such that

$$\frac{\partial U}{\partial x} = P, \quad \frac{\partial U}{\partial y} = Q.$$

Therefore (3.4) becomes $\frac{d}{dx}U(x, y(x)) = 0$, so that the solutions can be expressed by $U(x, y(x)) = C$. For the case when (3.5) is not satisfied, Clairaut and Euler investigated the possibility of multiplying (3.4) by a suitable factor $M(x, y)$, which sometimes allows the equation $MP + MQy' = 0$ to satisfy (3.5).

Second Order Equations

Even more than for first order equations, the solution of *second* order equations by integration is very seldom possible. Besides linear equations with constant coefficients, whose solutions for the second order case were already known to Newton, several tricks of reduction are possible, as for example the following:

For a *linear equation*

$$y'' = a(x)y' + b(x)y$$

we make the substitution (Riccati 1723, Euler 1728)

$$y = \exp\left(\int p(x) dx\right). \quad (3.6)$$

The derivatives of this function contain only derivatives of p of lower order

$$y' = p \cdot \exp\left(\int p(x) dx\right), \quad y'' = (p^2 + p') \cdot \exp\left(\int p(x) dx\right)$$

so that inserting this into the differential equation, after division by y , leads to a *lower order* equation

$$p^2 + p' = a(x)p + b(x) \quad (3.7)$$

which, however, is nonlinear.

If the equation is *independent* of y , $y'' = f(x, y')$, it is natural to put $y' = v$ which gives $v' = f(x, v)$.

An important case is that of *equations independent of x* :

$$y'' = f(y, y').$$

Here we consider y' as function of y : $y' = p(y)$. Then the chain rule gives $y'' = p'p = f(y, p)$, which is a first order equation. When the function $p(y)$ has been found, it remains to integrate $y' = p(y)$, which is an equation of type (3.1) (Riccati (1712): “Per liberare la premessa formula dalle seconde differenze, . . . , chiamo p la sunnormale BF . . . ”, see also Euler (1769), Problema 96, p. 33).

The investigation of all possible differential equations which can be integrated by analytical methods was begun by Euler. His results have been collected, in

more than 800 pages, in Volumes XXII and XXIII of Euler's Opera Omnia. For a more recent discussion see Ince (1944), p. 16-61. An irreplaceable document on this subject is the book of Kamke (1942). It contains, besides a description of the solution methods and general properties of the solutions, a systematically ordered list of more than 1500 differential equations with their solutions and references to the literature.

The computations, even for very simple looking equations, soon become very complicated and one quickly began to understand that elementary solutions would not always be possible. It was Liouville (1841) who gave the first *proof* of the fact that certain equations, such as $y' = x^2 + y^2$, cannot be solved in terms of elementary functions. Therefore, in the 19th century mathematicians became more and more interested in general existence theorems and in numerical methods for the computation of the solutions.

Exercises

1. Solve Newton's equation (2.1) by quadrature.
2. Solve Leibniz' equation (2.3) in terms of elementary functions.
Hint. The integral for y might cause trouble. Use the substitution $a^2 - y^2 = u^2$, $-ydy = udu$.
3. Solve and draw the solutions of $y' = f(y)$ where $f(y) = \sqrt{|y|}$.
4. Solve the master-and-dog problem: a dog runs with speed w in the direction of his master, who walks with speed v along the y -axis. This leads to the differential equation

$$(xy')' = -\frac{v}{w} \sqrt{1 + (y')^2}.$$
5. Solve the equation $my'' = -k/y^2$, which describes a body falling according to Newton's law of gravitation.
6. Verify that the cycloid

$$x - x_0 = R(\tau - \sin \tau), \quad y - h = R(1 - \cos \tau), \quad R = \frac{1}{4gK^2}$$

satisfies the differential equation (2.4) for the brachistochrone problem. Solving (2.4) in a forward manner, one arrives after some simplifications at the integral

$$\int \sqrt{\frac{y}{1-y}} dy,$$

which is computed by the substitution $y = (\sin t)^2$.

7. Reduce the “Bernoulli equation” (Jac. Bernoulli 1695)

$$y' + f(x)y = g(x)y^n$$

with the help of the coordinate transformation $z(x) = (y(x))^q$ and a suitable choice of q , to a linear equation (Leibniz, Acta Erud. 1696, p. 145, Joh. Bernoulli, Acta Erud. 1697, p. 113).

8. Compute the “Linea Catenaria” of the hanging rope. The solution was given by Joh. Bernoulli (1691) and Leibniz (1691) (see Fig. 3.2) without any hint.

Hint. (Joh. Bernoulli, “Lectiones ... in usum III. Marchionis Hospitalii” 1691/92). Let H resp. V be the horizontal resp. vertical component of the tension in the rope (Fig. 3.1). Then $H = a$ is a constant and $V = q \cdot s$ is proportional to the arc length. This leads to $Cp = s$ or $Cdp = ds$ i.e., $Cdp = \sqrt{1 + p^2}dx$, where $p = y'$, a differential equation.

Result. $y = K + C \cosh\left(\frac{x - x_0}{C}\right)$.

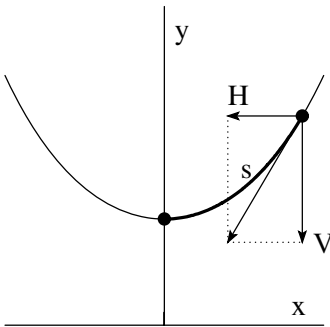


Fig. 3.1. Solution of the Catenary problem

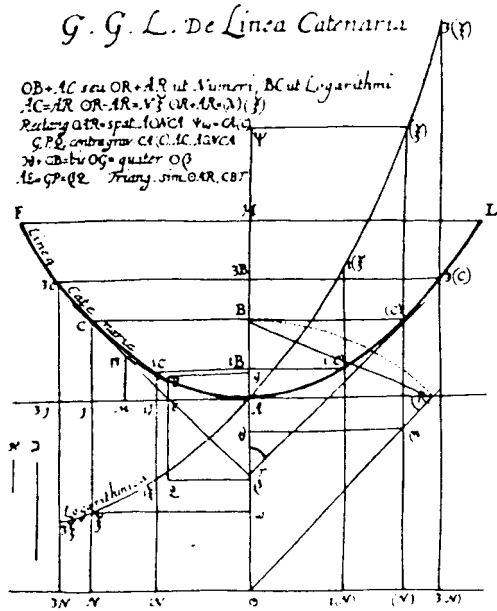


Fig. 3.2. “Linea Catenaria” drawn by Leibniz (1691)

I.4 Linear Differential Equations

Lisez Euler, lisez Euler, c'est notre maître à tous. (Laplace)

[Euler] ... c'est un homme peu amusant, mais un très-grand Géomètre. (D'Alembert, letter to Voltaire, March 3, 1766)

[Euler] ... un Géomètre borgne, dont les oreilles ne sont pas faites pour sentir les délicatesses de la poésie. (Frédéric II, in a letter to Voltaire)

Following in the footsteps of Euler (1743), we want to understand the general solution of n th order linear differential equations. We say that the equation

$$\mathcal{L}(y) := a_n(x)y^{(n)} + a_{n-1}(x)y^{(n-1)} + \dots + a_0(x)y = 0 \quad (4.1)$$

with given functions $a_0(x), \dots, a_n(x)$ is *homogeneous*. If n solutions $u_1(x), \dots, u_n(x)$ of (4.1) are known, then any linear combination

$$y(x) = C_1 u_1(x) + \dots + C_n u_n(x) \quad (4.2)$$

with constant coefficients C_1, \dots, C_n is also a solution of (4.1), since all derivatives of y appear only linearly in (4.1).

Equations with Constant Coefficients

Let us first consider the special case

$$y^{(n)}(x) = 0. \quad (4.3)$$

This can be integrated once to give $y^{(n-1)}(x) = C_1$, then $y^{(n-2)}(x) = C_1 x + C_2$, etc. Replacing at the end the arbitrary constants C_i by new ones, we finally obtain

$$y(x) = C_1 x^{n-1} + C_2 x^{n-2} + \dots + C_n.$$

Thus there are n “free parameters” in the “general solution” of (4.3). Euler’s intuition, after some more examples, also expected the same result for the general equation (4.1). This fact, however, only became completely clear many years later.

We now treat the general equation with constant coefficients,

$$y^{(n)} + A_{n-1}y^{(n-1)} + \dots + A_0y = 0. \quad (4.4)$$

Our problem is to find a basis of n linearly independent solutions $u_1(x), \dots, u_n(x)$. To this end, Euler’s inspiration was guided by the transformation (3.6), (3.7) above: if $a(x)$ and $b(x)$ are constants, we assume p constant in (3.7) so that p' vanishes, and we obtain the quadratic equation $p^2 = ap + b$. For any root of this

equation, (3.6) then becomes $y = e^{px}$. In the general case we thus assume $y = e^{px}$ with an unknown constant p , so that (4.4) leads to the *characteristic equation*

$$p^n + A_{n-1}p^{n-1} + \dots + A_0 = 0. \quad (4.5)$$

If the roots p_1, \dots, p_n of equation (4.5) are distinct, all solutions of (4.4) are given by

$$y(x) = C_1 e^{p_1 x} + \dots + C_n e^{p_n x}. \quad (4.6)$$

It is curious to see that the “brightest mathematicians of the world” struggled for many decades to find this solution, which appears so trivial to today’s students.

A difficulty arises with the solution (4.6) when (4.5) does not possess n distinct roots. Consider, with Euler, the example

$$y'' - 2qy' + q^2y = 0. \quad (4.7)$$

Here $p = q$ is a double root of the corresponding characteristic equation. If we set

$$y = e^{qx}u, \quad (4.8)$$

(4.7) becomes $u'' = 0$, which brings us back to (4.3). So the general solution of (4.7) is given by $y(x) = e^{qx}(C_1x + C_2)$ (see also Exercise 5 below). After some more examples of this type, one sees that the transformation (4.8) effects a *shift* of the characteristic polynomial, so that if q is a root of multiplicity k , we obtain for u an equation ending with $\dots + Bu^{(k+1)} + Cu^{(k)} = 0$. Therefore

$$e^{qx}(C_1x^{k-1} + \dots + C_k)$$

gives us k independent solutions.

Finally, for a pair of complex roots $p = \alpha \pm i\beta$ the solutions $e^{(\alpha+i\beta)x}$, $e^{(\alpha-i\beta)x}$ can be replaced by the real functions

$$e^{\alpha x}(C_1 \cos \beta x + C_2 \sin \beta x).$$

The study of the *inhomogeneous* equation

$$\mathcal{L}(y) = f(x) \quad (4.9)$$

was begun in Euler (1750), p. 13. We mention from this work the case where $f(x)$ is a polynomial, say for example the equation

$$Ay'' + By' + Cy = ax^2 + bx + c. \quad (4.10)$$

Here Euler puts $y(x) = Ex^2 + Fx + G + v(x)$. Inserting this into (4.10) and eliminating all possible powers of x , one obtains

$$\begin{aligned} CE = a, \quad CF + 2BE = b, \quad CG + BF + 2AE = c, \\ Av'' + Bv' + Cv = 0. \end{aligned}$$

This allows us, when C is different from zero, to compute E, F and G and we observe that *the general solution of the inhomogeneous equation is the sum of a*

particular solution of it and of the general solution of the corresponding homogeneous equation. This is also true in the general case and can be verified by trivial linear algebra.

The above method of searching for a particular solution with the help of unknown coefficients works similarly if $f(x)$ is composed of exponential, sine, or cosine functions and is often called the “fast method”. We see with pleasure that it was historically the first method to be discovered.

Variation of Constants

The general treatment of the inhomogeneous equation

$$a_n(x)y^{(n)} + \dots + a_0(x)y = f(x) \quad (4.11)$$

is due to Lagrange (1775) (“... par une nouvelle méthode aussi simple qu’on puisse le désirer”, see also Lagrange (1788), seconde partie, Sec. V.) We assume known n independent solutions $u_1(x), \dots, u_n(x)$ of the *homogeneous* equation. We then set, in extension of the method employed for (3.2), instead of (4.2)

$$y(x) = c_1(x)u_1(x) + \dots + c_n(x)u_n(x) \quad (4.12)$$

with unknown functions $c_i(x)$ (“method of variation of constants”). We have to insert (4.12) into (4.11) and thus compute the first derivative

$$y' = \sum_{i=1}^n c'_i u_i + \sum_{i=1}^n c_i u'_i.$$

If we continue blindly to differentiate in this way, we soon obtain complicated and useless formulas. Therefore Lagrange astutely requires the first term to vanish and puts

$$\sum_{i=1}^n c'_i u_i^{(j)} = 0 \quad j = 0, \quad \text{then also for } j = 1, \dots, n-2. \quad (4.13)$$

Then repeated differentiation of y , with continued elimination of the undesired terms (4.13), gives

$$\begin{aligned} y' &= \sum_{i=1}^n c_i u'_i, & \dots & & y^{(n-1)} &= \sum_{i=1}^n c_i u_i^{(n-1)}, \\ y^{(n)} &= \sum_{i=1}^n c'_i u_i^{(n-1)} + \sum_{i=1}^n c_i u_i^{(n)}. \end{aligned}$$

If we insert this into (4.11), we observe wonderful cancellations due to the fact that the $u_i(x)$ satisfy the homogeneous equation, and finally obtain, together with (4.13),

$$\begin{pmatrix} u_1 & \cdots & u_n \\ u_1' & \cdots & u_n' \\ \vdots & & \vdots \\ u_1^{(n-1)} & \cdots & u_n^{(n-1)} \end{pmatrix} \begin{pmatrix} c_1' \\ c_2' \\ \vdots \\ c_n' \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ f(x)/a_n(x) \end{pmatrix}. \quad (4.14)$$

This is a linear system, whose determinant is called the “Wronskian” and whose solution yields $c_1'(x), \dots, c_n'(x)$ and after integration $c_1(x), \dots, c_n(x)$.

Much more insight into this formula will be possible in Section I.11.

Exercises

1. Find the solution “huius aequationis differentialis quarti gradus” $a^4 y^{(4)} + y = 0$, $a^4 y^{(4)} - y = 0$; solve the equation “septimi gradus” $y^{(7)} + y^{(5)} + y^{(4)} + y^{(3)} + y^{(2)} + y = 0$. (Euler 1743, Ex. 4, 5, 6).

2. Solve by Euler’s technique $y'' - 3y' - 4y = \cos x$ and $y'' + y = \cos x$.

Hint. In the first case the particular solution can be searched for in the form $E \cos x + F \sin x$. In the second case (which corresponds to a resonance in the equation) one puts $Ex \cos x + Fx \sin x$ just as in the solution of (4.7).

3. Find the solution of

$$y'' - 3y' - 4y = g(x), \quad g(x) = \begin{cases} \cos(x) & 0 \leq x \leq \pi/2 \\ 0 & \pi/2 \leq x \end{cases}$$

such that $y(0) = y'(0) = 0$,

a) by using the solution of Exercise 2,

b) by the method of Lagrange (variation of constants).

4. (Reduction of the order if one solution is known). Suppose that a nonzero solution $u_1(x)$ of $y'' + a_1(x)y' + a_0(x)y = 0$ is known. Show that a second independent solution can be found by putting $u_2(x) = c(x)u_1(x)$.

5. Treat the case of multiple characteristic values (4.7) by considering them as a limiting case $p_2 \rightarrow p_1$ and using the solutions

$$u_1(x) = e^{p_1 x}, \quad u_2(x) = \lim_{p_2 \rightarrow p_1} \frac{e^{p_2 x} - e^{p_1 x}}{p_2 - p_1} = \frac{\partial e^{p_1 x}}{\partial p_1}, \text{ etc.}$$

(d’Alembert (1748), p. 284: “Enfin, si les valeurs de p & de p' sont égales, au lieu de les supposer telles, on supposera $p = a + \alpha$, $p' = a - \alpha$, α étant quantité infiniment petite . . .”).

I.5 Equations with Weak Singularities

Der Mathematiker weiss sich ohnedies beim Auftreten von singulären Stellen gegebenenfalls leicht zu helfen. (K. Heun 1900)

Many equations occurring in applications possess *singularities*, i.e., points at which the function $f(x, y)$ of the differential equation becomes infinite. We study in some detail the classical treatment of such equations, since numerical methods, which will be discussed later in this book, often fail at the singular point, at least if they are not applied carefully.

Linear Equations

As a first example, consider the equation

$$y' = \frac{q + bx}{x} y, \quad q \neq 0 \quad (5.1)$$

with a singularity at $x = 0$. Its solution, using the method of separation of variables (3.1), is

$$y(x) = Cx^q e^{bx} = C(x^q + bx^{q+1} + \dots). \quad (5.2)$$

These solutions are plotted in Fig. 5.1 for different values of q and show the fundamental difference in the behaviour of the solutions in dependence of q .

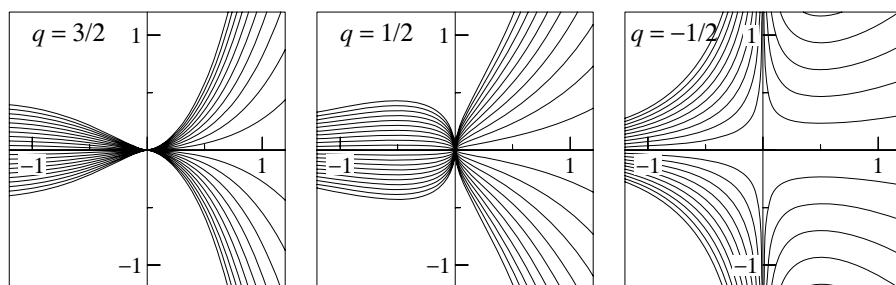


Fig. 5.1. Solutions of (5.1) for $b = 1$

Euler started a systematic study of equations with singularities. He asked which type of equation of the second order can conveniently be solved by a series as in (5.2) (Euler 1769, Problema 122, p. 177, “. . . quas commode per series resolvere licet. . .”). He found the equation

$$\mathcal{L}y := x^2(a + bx)y'' + x(c + ex)y' + (f + gx)y = 0. \quad (5.3)$$

Let us put $y = x^q(A_0 + A_1x + A_2x^2 + \dots)$ with $A_0 \neq 0$ and insert this into (5.3). We observe that the powers x^2 and x which are multiplied by y'' and y' , respectively, just re-establish what has been lost by the differentiations and obtain by comparing equal powers of x

$$(q(q-1)a + qc + f)A_0 = 0 \quad (5.4a)$$

$$\begin{aligned} ((q+i)(q+i-1)a + (q+i)c + f)A_i \\ = -((q+i-1)(q+i-2)b + (q+i-1)e + g)A_{i-1} \end{aligned} \quad (5.4b)$$

for $i = 1, 2, 3, \dots$. In order to get $A_0 \neq 0$, q has to be a root of the *index equation*

$$\chi(q) := q(q-1)a + qc + f = 0. \quad (5.5)$$

For $a \neq 0$ there are two characteristic roots q_1 and q_2 of (5.5). Since the left-hand side of (5.4b) is of the form $\chi(q+i)A_i = \dots$, this relation allows us to compute A_1, A_2, A_3, \dots at least for q_1 (if the roots are ordered such that $\operatorname{Re} q_1 \geq \operatorname{Re} q_2$). Thus we have obtained a first non-zero solution of (5.3). A second linearly independent solution for $q = q_2$ is obtained in the same way if $q_1 - q_2$ is not an integer.

Case of double roots. Euler found a second solution in this case with the inspiration of some acrobatic heuristics (Euler 1769, p. 150: “. . . quod $\frac{x^0}{0}$ aequivalet ipsi $\ell x x \dots$ ”). Fuchs (1866, 1868) then wrote a monumental paper on the form of all solutions for the general equation of order n , based on complicated calculations. A very elegant idea was then found by Frobenius (1873): fix A_0 , say as $A_0(q) = 1$, completely ignore the index equation, choose q arbitrarily and consider the coefficients of the recursion (5.4b) as functions of q to obtain the series

$$y(x, q) = x^q \sum_{i=0}^{\infty} A_i(q)x^i, \quad (5.6)$$

whose convergence is discussed in Exercise 8 below. Since all conditions (5.4b) are satisfied, with the exception of (5.4a), we have

$$\mathcal{L}y(x, q) = \chi(q)x^q. \quad (5.7)$$

A second independent solution is now found simply by differentiating (5.7) with respect to q :

$$\mathcal{L}\left(\frac{\partial y}{\partial q}(x, q)\right) = \chi(q) \cdot \log x \cdot x^q + \chi'(q) \cdot x^q. \quad (5.8)$$

If we set $q = q_1$

$$\frac{\partial y}{\partial q}(x, q_1) = \log x \cdot y(x, q_1) + x^{q_1} \sum_{i=0}^{\infty} A'_i(q_1) x^i, \quad (5.9)$$

we obtain the desired second solution since $\chi(q_1) = \chi'(q_1) = 0$ (remember that q_1 is a double root of χ).

The case $q_1 - q_2 = m \in \mathbb{Z}$, $m \geq 1$. In this case we define a function $z(x)$ by satisfying $A_0(q) = 1$ and the recursion (5.4b) for all i with the exception of $i = m$. Then

$$\mathcal{L}z = \chi(q)x^q + Cx^{q+m} \quad (5.10)$$

where C is some constant. For $q = q_2$ the first term in (5.10) vanishes and a comparison with (5.8) shows that

$$\chi'(q_1)z(x) - C \frac{\partial y}{\partial q}(x, q_1) \quad (5.11)$$

is the required second solution of (5.3).

Euler (1778) later remarked that the formulas obtained become particularly elegant, if one starts from the differential equation

$$x(1-x)y'' + (c - (a+b+1)x)y' - aby = 0 \quad (5.12)$$

instead of from (5.3). Here, the above method leads to

$$q(q-1) + cq = 0, \quad q_1 = 0, \quad q_2 = 1 - c, \quad (5.13)$$

$$A_{i+1} = \frac{(a+i)(b+i)}{(c+i)(1+i)} A_i \quad \text{for } q_1 = 0. \quad (5.14)$$

The resulting solutions, later named *hypergeometric functions*, became particularly famous throughout the 19th century with the work of Gauss (1812).

More generally, the above method works in the case of a differential equation

$$x^2 y'' + xa(x)y' + b(x)y = 0 \quad (5.15)$$

where $a(x)$ and $b(x)$ are regular analytic functions. One then says that 0 is a *regular singular point*. Similarly, we say that the equation $(x - x_0)^2 y'' + (x - x_0)a(x)y' + b(x)y = 0$ possesses the regular singular point x_0 . In this case solutions can be obtained by the use of algebraic singularities $(x - x_0)^q$.

Finally, we also want to study the behaviour at *infinity* for an equation of the form

$$a(x)y'' + b(x)y' + c(x)y = 0. \quad (5.16a)$$

For this, we use the coordinate transformation $t = 1/x$, $z(t) = y(x)$ which yields

$$t^4 a\left(\frac{1}{t}\right) z'' + \left(2t^3 a\left(\frac{1}{t}\right) - t^2 b\left(\frac{1}{t}\right)\right) z' + c\left(\frac{1}{t}\right) z = 0. \quad (5.16b)$$

∞ is called a regular singular point of (5.16a) if 0 is a regular singular point of (5.16b). For examples see Exercise 9.

Nonlinear Equations

For nonlinear equations also, the above method sometimes allows one to obtain, if not the complete series of the solution, at least a couple of terms.

EXEMPLUM. Let us see what happens if we try to solve the classical brachystochrone problem (2.4) by a series. We suppose $h = 0$ and the initial value $y(0) = 0$. We write the equation as

$$(y')^2 = \frac{L}{y} - 1 \quad \text{or} \quad y(y')^2 + y = L. \quad (5.17)$$

At the initial point $y(0) = 0$, y' becomes infinite and most numerical methods would fail. We search for a solution of the form $y = A_0 x^q$. This gives in (5.17) $q^2 A_0^3 x^{3q-2} + A_0 x^q = L$. Due to the initial value we have that $y(x)$ becomes negligible for small values of x . We thus set the first term equal to L and obtain $3q - 2 = 0$ and $q^2 A_0^3 = L$. So

$$u(x) = \left(\frac{9Lx^2}{4} \right)^{1/3} \quad (5.18)$$

is a first approximate solution. The idea is now to use (5.18) just to escape from the initial point with a small x , and then to continue the solution with any numerical step-by-step procedure from the later chapters.

A more refined approximation could be tried in the form $y = A_0 x^q + A_1 x^{q+r}$. This gives with (5.17)

$$q^2 A_0^3 x^{3q-2} + q(3q+2r)A_0^2 A_1 x^{3q+r-2} + A_0 x^q + \dots = L.$$

We use the second term to neutralize the third one, which gives $3q + r - 2 = q$ or $r = q = 2/3$ and $5q^2 A_0 A_1 = -1$. Therefore

$$v(x) = \left(\frac{9Lx^2}{4} \right)^{1/3} - \left(\frac{9^2 x^4}{4^2 L 5^3} \right)^{1/3} \quad (5.19)$$

is a better approximation. The following numerical results illustrate the utility of the approximations (5.18) and (5.19) compared with the correct solution $y(x)$ from I.3, Exercise 6, with $L = 2$:

$x = 0.10$	$y(x) = 0.342839$	$u(x) = 0.355689$	$v(x) = 0.343038$
$x = 0.01$	$y(x) = 0.076042$	$u(x) = 0.076631$	$v(x) = 0.076044.$

Exercises

1. Compute the general solution of the equation $x^2 y'' + xy' + gx^n y = 0$ with g constant (Euler 1769, Problema 123, Exemplum 1).

2. Apply the technique of Euler to the *Bessel equation*

$$x^2 y'' + xy' + (x^2 - g^2)y = 0.$$

Sketch the solutions obtained for $g = 2/3$ and $g = 10/3$.

3. Compute the solutions of the equations

$$x^2 y'' - 2xy' + y = 0 \quad \text{and} \quad x^2 y'' - 3xy' + 4y = 0.$$

Equations of this type are often called Euler's or even Cauchy's equation. Its solution, however, was already known to Joh. Bernoulli.

4. (Euler 1769, Probl. 123, Exempl. 2). Let

$$y(x) = \int_0^{2\pi} \sqrt{\sin^2 s + x^2 \cos^2 s} \, ds$$

be the perimeter of the ellipse with axes 1 and $x < 1$.

- a) Verify that $y(x)$ satisfies the differential equation

$$x(1 - x^2)y'' - (1 + x^2)y' + xy = 0. \quad (5.20)$$

- b) Compute the solutions of this equation.

- c) Show that the coordinate change $x^2 = t$, $y(x) = z(t)$ transforms (5.20) to a hypergeometric equation (5.12).

Hint. The computations for a) lead to the integral

$$\int_0^{2\pi} \frac{1 - 2 \cos^2 s + q^2 \cos^4 s}{(1 - q^2 \cos^2 s)^{3/2}} \, ds, \quad q^2 = 1 - x^2$$

which must be shown to be zero. Develop this into a power series in q^2 .

5. Try to solve the equation

$$x^2 y'' + (3x - 1)y' + y = 0$$

with the help of a series (5.6) and study its convergence.

6. Find a series of the type

$$y = A_0 x^q + A_1 x^{q+s} + A_2 x^{q+2s} + \dots$$

which solves the nonlinear "Emden-Fowler equation" of astrophysics $(x^2 y')' + y^2 x^{-1/2} = 0$ in the neighbourhood of $x = 0$.

7. Approximate the solution of Leibniz's equation (2.3) in the neighbourhood of the singular initial value $y(0) = a$ by a function of the type $y(x) = a - Cx^q$. Compare the result with the correct solution of Exercise 2 of I.3.
8. Show that the radius of convergence of series (5.6) is given by
- $$\text{i) } r = |a/b| \qquad \text{ii) } r = 1$$
- for the coefficients given by (5.4) and (5.14), respectively.
9. Show that the point ∞ is a regular singular point for the hypergeometric equation (5.12), but not for the Bessel equation of Exercise 2.
10. Consider the initial value problem

$$y' = \frac{\lambda}{x} y + g(x), \quad y(0) = 0. \quad (5.21)$$

- a) Prove that if $\lambda \leq 0$, the problem (5.21) possesses a unique solution for $x \geq 0$;
- b) If $g(x)$ is k -times differentiable and $\lambda \leq 0$, then the solution $y(x)$ is $(k+1)$ -times differentiable for $x \geq 0$ and we have

$$y^{(j)}(0) = \left(1 - \frac{\lambda}{j}\right)^{-1} g^{(j-1)}(0), \quad j = 1, 2, \dots$$

I.6 Systems of Equations

En général on peut supposer que l'Equation différentio-différentielle de la Courbe ADE est $\varphi dt^2 = \pm dde \dots$ (d'Alembert 1743, p. 16)

Parmi tant de chefs-d'œuvre que l'on doit à son génie [de Lagrange], sa *Mécanique* est sans contredit le plus grand, le plus remarquable et le plus important. (M. Delambre, Oeuvres de Lagrange, vol. 1, p. XLIX)

Newton (1687) distilled from the known solutions of planetary motion (the Kepler laws) his “Lex secunda” together with the universal law of gravitation. It was mainly the “Dynamique” of d'Alembert (1743) which introduced, the other way round, second order differential equations as a general tool for computing mechanical motion. Thus, Euler (1747) studied the movement of planets via the equations in 3-space

$$m \frac{d^2x}{dt^2} = X, \quad m \frac{d^2y}{dt^2} = Y, \quad m \frac{d^2z}{dt^2} = Z, \quad (6.1)$$

where X, Y, Z are the forces in the three directions. (“... & par ce moyen j'évite quantité de recherches pénibles”).

The Vibrating String and Propagation of Sound

Suppose a string is represented by a sequence of identical and equidistant mass points and denote by $y_1(t), y_2(t), \dots$ the deviation of these mass points from the equilibrium position (Fig. 6.1a). If the deviations are supposed small (“fort petites”), the repelling force for the i -th mass point is proportional to $-y_{i-1} + 2y_i - y_{i+1}$ (Brook Taylor 1715, Johann Bernoulli 1727). Therefore equations (6.1) become

$$\begin{aligned} y_1'' &= K^2(-2y_1 + y_2) \\ y_2'' &= K^2(y_1 - 2y_2 + y_3) \\ &\dots \\ y_n'' &= K^2(y_{n-1} - 2y_n). \end{aligned} \quad (6.2)$$

This is a system of n linear differential equations. Since the finite differences $y_{i-1} - 2y_i + y_{i+1} \approx c^2 \frac{\partial^2 y}{\partial x^2}$, equation (6.2) becomes, by the “inverse” method of lines, the famous partial differential equation (d'Alembert 1747)

$$\frac{\partial^2 u}{\partial t^2} = a^2 \frac{\partial^2 u}{\partial x^2}$$

for the vibrating string.

The *propagation of sound* is modelled similarly (Lagrange 1759): we suppose the medium to be a sequence of mass points and denote by $y_1(t)$, $y_2(t)$, \dots their longitudinal displacements from the equilibrium position (see Fig. 6.1b). Then by Hooke's law of elasticity the repelling forces are proportional to the differences of displacements $(y_{i-1} - y_i) - (y_i - y_{i+1})$. This leads to equations (6.2) again ("En examinant les équations, \dots je me suis bientôt aperçu qu'elles ne différaient nullement de celles qui appartiennent au problème de *chordis vibrantibus* \dots ").

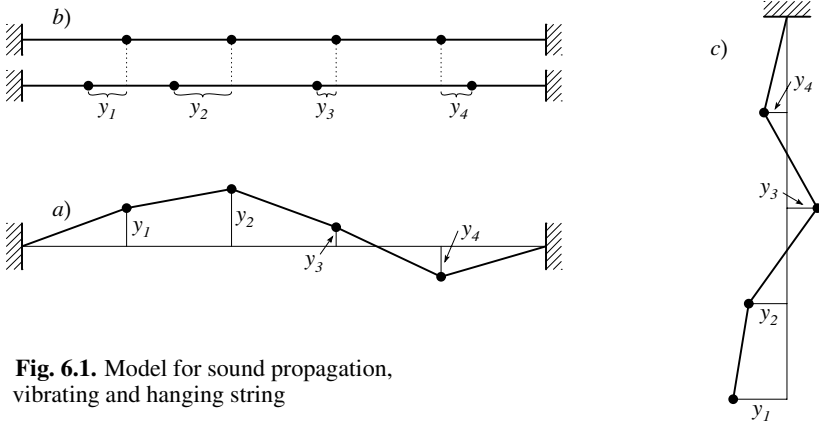


Fig. 6.1. Model for sound propagation, vibrating and hanging string

Another example, treated by Daniel Bernoulli (1732) and by Lagrange (1762, Nr. 36), is that of mass points attached to a *hanging* string (Fig. 6.1c). Here the tension in the string becomes greater in the upper part of the string and we have the following equations of movement

$$\begin{aligned} y_1'' &= K^2(-y_1 + y_2) \\ y_2'' &= K^2(y_1 - 3y_2 + 2y_3) \\ y_3'' &= K^2(2y_2 - 5y_3 + 3y_4) \\ &\dots \\ y_n'' &= K^2((n-1)y_{n-1} - (2n-1)y_n). \end{aligned} \quad (6.3)$$

In all these examples, of course, the deviations y_i are supposed to be "infinitely" small, so that linear models are realistic.

Using a notation which came into use only a century later, we write these equations in the form

$$y_i'' = \sum_{j=1}^n a_{ij}y_j, \quad i = 1, \dots, n, \quad (6.4)$$

which is a system of 2nd order linear equations with constant coefficients. La-

grange solves system (6.4) by putting $y_i = c_i e^{pt}$, which leads to

$$p^2 c_i = \sum_{j=1}^n a_{ij} c_j, \quad i = 1, \dots, n \quad (6.5)$$

so that p^2 must be an *eigenvalue* of the matrix $A = (a_{ij})$ and $c = (c_1, \dots, c_n)^T$ a corresponding *eigenvector*. We see here the first appearance of an eigenvalue problem.

Lagrange (1762, Nr. 30) then explains that the equations (6.5) are solved by computing $c_2/c_1, \dots, c_n/c_1$ as functions of p from $n-1$ equations and by inserting these results into the last equation. This leads to a polynomial of degree n (in fact, the *characteristic polynomial*) to obtain n different roots for p^2 . We thus get $2n$ solutions $y_i^{(j)} = c_i^{(j)} \exp(\pm p_j t)$ and the general solution as linear combinations of these.

A complication arises when the characteristic polynomial possesses *multiple roots*. In this case, Lagrange (in his famous “*Mécanique Analytique*” of 1788, seconde partie, sect.VI, No.7) affirms the presence of “secular” terms similar to the formulas following (4.8). This, however, is not completely true, as became clear only a century later (see e.g., Weierstrass (1858), p.243: “. . . um bei dieser Gelegenheit einen Irrtum zu berichtigen, der sich in der Lagrange’schen Theorie der kleinen Schwingungen, sowie in allen späteren mir bekannten Darstellungen derselben, findet.”). We therefore postpone this subject to Section I.12.

We solve equations (6.2) in detail, since the results obtained are of particular importance (Lagrange 1759). The corresponding eigenvalue problem (6.5) becomes in this case $p^2 c_1 = K^2(-2c_1 + c_2)$, $p^2 c_i = K^2(c_{i-1} - 2c_i + c_{i+1})$ for $i = 2, \dots, n-1$ and $p^2 c_n = K^2(c_{n-1} - 2c_n)$. We introduce $p^2/K^2 + 2 = q$, so that

$$c_{j+1} - qc_j + c_{j-1} = 0, \quad c_0 = 0, \quad c_{n+1} = 0. \quad (6.6)$$

This means that the c_i are the solutions of a *difference equation* and therefore $c_j = Aa^j + Bb^j$ where a and b are the roots of the corresponding characteristic equation $z^2 - qz + 1 = 0$, hence

$$a + b = q, \quad ab = 1.$$

The condition $c_0 = 0$ of (6.6), which means that $A + B = 0$, shows that $c_j = A(a^j - b^j)$ with $A \neq 0$. The second condition $c_{n+1} = 0$, or equivalently $(a/b)^{n+1} = 1$, implies together with $ab = 1$ that

$$a = \exp\left(\frac{k\pi i}{n+1}\right), \quad b = \exp\left(\frac{-k\pi i}{n+1}\right)$$

for some $k = 1, \dots, n$. Thus we obtain

$$q_k = 2 \cos \frac{\pi k}{n+1}, \quad k = 1, \dots, n, \quad (6.7a)$$

$$p_k^2 = 2K^2 \left(\cos \frac{\pi k}{n+1} - 1 \right) = -4K^2 \left(\sin \frac{\pi k}{2n+2} \right)^2. \quad (6.7b)$$

Finally, Euler's formula from 1740, $e^{ix} - e^{-ix} = 2i \sin x$ ("... si familière aujourd'hui aux Géomètres") gives for the eigenvectors (with $A = -i/2$)

$$c_j^{(k)} = \sin \frac{jk\pi}{n+1}, \quad j, k = 1, \dots, n. \quad (6.8)$$

Since the p_k are purely imaginary, we also use for $\exp(\pm p_k t)$ the "familière" formula and obtain the general solution

$$y_j(t) = \sum_{k=1}^n \sin \frac{jk\pi}{n+1} (a_k \cos r_k t + b_k \sin r_k t), \quad r_k = 2K \sin \frac{\pi k}{2n+2}. \quad (6.9)$$

Lagrange then observed after some lengthy calculations, which are today seen by using the orthogonality relations

$$\sum_{\ell=1}^n \sin \frac{\ell j \pi}{n+1} \sin \frac{\ell k \pi}{n+1} = \begin{cases} 0 & j \neq k \\ \frac{n+1}{2} & j = k \end{cases} \quad j, k = 1, \dots, n$$

that

$$a_k = \frac{2}{n+1} \sum_{j=1}^n \sin \frac{jk\pi}{n+1} y_j(0), \quad b_k = \frac{1}{r_k} \frac{2}{n+1} \sum_{j=1}^n \sin \frac{jk\pi}{n+1} y'_j(0)$$

are determined by the initial positions and velocities of the mass points. He also studied the case where n , the number of mass points, tends to infinity (so that, in the formula for r_k , $\sin x$ can be replaced by x) and stood, 50 years before Fourier, at the portal of Fourier series theory. "Mit welcher Gewandtheit, mit welchem Aufwande analytischer Kunstgriffe er auch den ersten Theil dieser Untersuchung durchführte, so liess der Uebergang vom Endlichen zum Unendlichen doch viel zu wünschen übrig. . ." (Riemann 1854).

Fourier

J'ajouterais que le livre de Fourier a une importance capitale dans l'histoire des mathématiques. (H. Poincaré 1893)

The first *first order systems* were motivated by the problem of heat conduction (Biot 1804, Fourier 1807). Fourier imagined a rod to be a sequence of molecules, whose temperatures we denote by y_i , and deduced from a law of Newton that the energy which a particle passes to its neighbours is proportional to the difference of their temperatures, i.e., $y_{i-1} - y_i$ to the left and $y_{i+1} - y_i$ to the right ("Lorsque deux molécules d'un même solide sont extrêmement voisines et ont des températures inégales, la molécule plus échauffée communique à celle qui l'est moins une quantité de chaleur exactement exprimée par le produit formé de la durée de l'instant,

de la différence extrêmement petite des températures, et d'une certaine fonction de la distance des molécules"). This long sentence means, in formulas, that the total gain of energy of the i th molecule is expressed by

$$y'_i = K^2(y_{i-1} - 2y_i + y_{i+1}), \quad (6.10)$$

or, in general by

$$y'_i = \sum_{j=1}^n a_{ij}y_j, \quad i = 1, \dots, n, \quad (6.11)$$

a first order system with constant coefficients.

By putting $y_i = c_i e^{pt}$, we now obtain the eigenvalue problem

$$pc_i = \sum_{j=1}^n a_{ij}c_j, \quad i = 1, \dots, n. \quad (6.12)$$

If we suppose the rod cooled to zero at both ends ($y_0 = y_{n+1} = 0$), we can use Lagrange eigenvectors from above and obtain the solution

$$y_j(t) = \sum_{k=1}^n a_k \sin \frac{jk\pi}{n+1} \exp(-w_k t), \quad w_k = 4K^2 \left(\sin \frac{\pi k}{2n+2} \right)^2. \quad (6.13)$$

By taking n larger and larger, Fourier arrived from (6.10) (again the inverse "method of lines") at his famous heat equation

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} \quad (6.14)$$

which was the origin of Fourier series theory.

Lagrangian Mechanics

Dies ist der kühne Weg, den *Lagrange* ..., freilich ohne ihn gehörig zu rechtfertigen, eingeschlagen hat.

(Jacobi 1842/43, Vorl. Dynamik, p. 13)

This combines d'Alembert's dynamics, the "principle of least action" of Leibniz–Maupertuis and the variational calculus; published in the monumental treatise "Mécanique Analytique" (1788). It furnishes an excellent means for obtaining the differential equations of motion for complicated mechanical systems (arbitrary coordinate systems, constraints, etc.).

If we define (with Poisson 1809) the "Lagrange function"

$$\mathcal{L} = T - U \quad (6.15)$$

where

$$T = m \frac{\dot{x}^2 + \dot{y}^2 + \dot{z}^2}{2} \quad (\text{kinetic energy}) \quad (6.16)$$

and U is the “potential energy” satisfying

$$\frac{\partial U}{\partial x} = -X, \quad \frac{\partial U}{\partial y} = -Y, \quad \frac{\partial U}{\partial z} = -Z \quad (6.17)$$

then the equations of motion (6.1) are *identical* to Euler’s equations (2.11) for the variational problem

$$\int_{t_0}^{t_1} \mathcal{L} \, dt = \min \quad (6.18)$$

(this, mainly through a misunderstanding of Jacobi, is often called “Hamilton’s principle”). The important idea is now to forget (6.16) and (6.17) and to apply (6.15) and (6.18) to *arbitrary mass points* and *arbitrary coordinate systems*.

Example. The *spherical pendulum* (Lagrange 1788, *Seconde partie*, Section VIII, Chap. II, §I). Let $\ell = 1$ and

$$\begin{aligned} x &= \sin \theta \cos \varphi \\ y &= \sin \theta \sin \varphi \\ z &= -\cos \theta. \end{aligned}$$

We set $m = g = 1$ and have

$$\begin{aligned} T &= \frac{1}{2} (\dot{x}^2 + \dot{y}^2 + \dot{z}^2) = \frac{1}{2} (\dot{\theta}^2 + \sin^2 \theta \cdot \dot{\varphi}^2) \\ U &= z = -\cos \theta \end{aligned} \quad (6.19)$$

so that (2.11) becomes

$$\begin{aligned} \mathcal{L}_\theta - \frac{d}{dt} (\mathcal{L}_{\dot{\theta}}) &= -\sin \theta + \sin \theta \cos \theta \cdot \dot{\varphi}^2 - \ddot{\theta} = 0 \\ \mathcal{L}_\varphi - \frac{d}{dt} (\mathcal{L}_{\dot{\varphi}}) &= -\sin^2 \theta \cdot \ddot{\varphi} - 2 \sin \theta \cos \theta \cdot \dot{\varphi} \cdot \dot{\theta} = 0. \end{aligned} \quad (6.20)$$

We have thus obtained, by simple calculus, the equations of motion for the problem. These equations cannot be solved analytically. A solution, computed numerically by a Runge-Kutta method (see Chapter II) is shown in Fig. 6.2.

In general, suppose that the mechanical system in question is described by n coordinates q_1, q_2, \dots, q_n and that $\mathcal{L} = T - U$ depends on $q_1, q_2, \dots, q_n, \dot{q}_1, \dot{q}_2, \dots, \dot{q}_n$. Then the equations of motion are

$$\frac{d}{dt} \mathcal{L}_{\dot{q}_i} = \sum_{k=1}^n \mathcal{L}_{\dot{q}_i \dot{q}_k} \ddot{q}_k + \sum_{k=1}^n \mathcal{L}_{\dot{q}_i q_k} \dot{q}_k = \mathcal{L}_{q_i}, \quad i = 1, \dots, n. \quad (6.21)$$

These equations allow several generalizations to time-dependent systems and non-conventional forces.

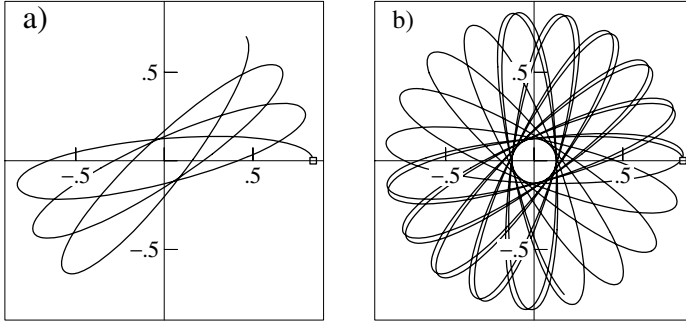


Fig. 6.2. Solution of the spherical pendulum, a) $0 \leq x \leq 20$, b) $0 \leq x \leq 100$
 $(\varphi_0 = 0, \quad \dot{\varphi}_0 = 0.17, \quad \theta_0 = 1, \quad \dot{\theta}_0 = 0)$

Hamiltonian Mechanics

Nach dem Erscheinen der ersten Ausgabe der *Mécanique analytique* wurde der wichtigste Fortschritt in der Umformung der Differentialgleichungen der Bewegung von *Poisson* ... gemacht ... im 15^{ten} Hefte des polytechnischen Journals ... Hier führt *Poisson* die Grössen $p = \partial T / \partial q'$... ein.
 (Jacobi 1842/43, Vorl. Dynamik, p. 67)

Hamilton, having worked for many years with variational principles (Fermat's principle) in his researches on optics, discovered at once that his ideas, after introducing a "principal function", allowed very elegant solutions for Kepler's motion of a planet (Hamilton 1833). He then undertook in several papers (Hamilton 1834, 1835) to revolutionize mechanics. After many pages of computation he thereby discovered that it was "more convenient in many respects" (Hamilton 1834, Math. Papers II, p. 161) to work with the momentum coordinates (idea of Poisson)

$$p_i = \frac{\partial \mathcal{L}}{\partial \dot{q}_i} \quad (6.22)$$

instead of \dot{q}_i , and with the function

$$H = \sum_{k=1}^n \dot{q}_k p_k - \mathcal{L} \quad (6.23)$$

considered as function of $q_1, \dots, q_n, p_1, \dots, p_n$. This idea, to let derivatives $\partial \mathcal{L} / \partial \dot{q}_i$ and independent variables p_i interchange their parts in order to simplify differential equations, is due to Legendre (1787). Differentiating (6.23) by the

chain rule, we obtain

$$\frac{\partial H}{\partial p_i} = \sum_{k=1}^n \frac{\partial \dot{q}_k}{\partial p_i} \cdot p_k + \dot{q}_i - \sum_{k=1}^n \frac{\partial \mathcal{L}}{\partial \dot{q}_k} \frac{\partial \dot{q}_k}{\partial p_i}$$

and

$$\frac{\partial H}{\partial q_i} = \sum_{k=1}^n \frac{\partial \dot{q}_k}{\partial q_i} \cdot p_k - \frac{\partial \mathcal{L}}{\partial q_i} - \sum_{k=1}^n \frac{\partial \mathcal{L}}{\partial \dot{q}_k} \frac{\partial \dot{q}_k}{\partial q_i}.$$

By (6.22) and (6.21) both formulas simplify to

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}, \quad i = 1, \dots, n. \quad (6.24)$$

These equations are marvellously symmetric “... and to integrate these differential equations of motion... is the chief and perhaps ultimately the only problem of mathematical dynamics” (Hamilton 1835). Jacobi (1843) called them *canonical* differential equations.

Remark. If the kinetic energy T is a quadratic function of the velocities \dot{q}_i , Euler’s identity (Euler 1755, Caput VII, § 224, “... si V fuerit functio homogenea...”) states that

$$2T = \sum_{k=1}^n \dot{q}_k \frac{\partial T}{\partial \dot{q}_k}. \quad (6.25)$$

If we further assume that the potential energy U is independent of \dot{q}_i , we obtain

$$H = \sum_{k=1}^n \dot{q}_k p_k - \mathcal{L} = \sum_{k=1}^n \dot{q}_k \frac{\partial T}{\partial \dot{q}_k} - \mathcal{L} = 2T - \mathcal{L} = T + U. \quad (6.26)$$

This is the *total* energy of the system.

Example. The spherical pendulum again. From (6.19) we have

$$p_\theta = \frac{\partial T}{\partial \dot{\theta}} = \dot{\theta}, \quad p_\varphi = \frac{\partial T}{\partial \dot{\varphi}} = \sin^2 \theta \cdot \dot{\varphi} \quad (6.27)$$

and, by eliminating the undesired variables $\dot{\theta}$ and $\dot{\varphi}$,

$$H = T + U = \frac{1}{2} \left(p_\theta^2 + \frac{p_\varphi^2}{\sin^2 \theta} \right) - \cos \theta. \quad (6.28)$$

Therefore (6.26) becomes

$$\begin{aligned} \dot{p}_\theta &= p_\varphi^2 \cdot \frac{\cos \theta}{\sin^3 \theta} - \sin \theta & \dot{p}_\varphi &= 0 \\ \dot{\theta} &= p_\theta & \dot{\varphi} &= \frac{p_\varphi}{\sin^2 \theta}. \end{aligned} \quad (6.29)$$

These equations appear to be a little simpler than Lagrange’s formulas (6.20). For example, we immediately see that $p_\varphi = \text{Const}$ (Kepler’s second law).

Exercises

1. Verify that, if $u(x)$ is sufficiently differentiable,

$$\frac{u(x-\delta) - 2u(x) + u(x+\delta)}{\delta^2} = u''(x) + \frac{\delta^2}{12} u^{(4)}(x) + \mathcal{O}(\delta^4).$$

Hint. Use Taylor series expansions for $u(x+\delta)$ and $u(x-\delta)$. This relation establishes the connection between (6.10) and (6.14) as well as between (6.2) and the wave equation.

2. Solve equation (6.3) for $n = 2$ and $n = 3$ by using the device of Lagrange described above (1762) and discover naturally the characteristic polynomial of the matrix.
3. Solve the first order system (6.11) with initial values $y_i(0) = (-1)^i$, where the matrix A is the same as in Exercise 2, and draw the solutions. Physically, this equation would represent a string with weights hanging, say, in honey.
4. Find the first terms of the development at the singular point $x = 0$ of the solutions of the following system of nonlinear equations

$$\begin{aligned} x^2 y'' + 2xy' &= 2yz^2 + \lambda x^2 y(y^2 - 1), & y(0) &= 0 \\ x^2 z'' &= z(z^2 - 1) + x^2 y^2 z, & z(0) &= 1 \end{aligned} \quad (6.30)$$

where λ is a constant parameter. Equations (6.30) are the Euler equations for the variational problem

$$I = \int_0^\infty \left((z')^2 + \frac{x^2 (y')^2}{2} + \frac{(z^2 - 1)^2}{2x^2} + y^2 z^2 + \frac{\lambda}{4} x^2 (y^2 - 1)^2 \right) dx,$$

$$y(\infty) = 1, \quad z(\infty) = 0$$

which gives the mass of a “monopole” in nuclear physics (see 't Hooft 1974).

5. Prove that the Hamiltonian function $H(q_1, \dots, q_n, p_1, \dots, p_n)$ is a first integral for the system (6.24), i.e., every solution satisfies

$$H(q_1(t), \dots, q_n(t), p_1(t), \dots, p_n(t)) = \text{Const.}$$

I.7 A General Existence Theorem

M. Cauchy annonce, que, pour se conformer au voeu du Conseil, il ne s'attachera plus à donner, comme il a fait jusqu'à présent, des démonstrations parfaitement rigoureuses.

(Conseil d'instruction de l'Ecole polytechnique, 24 nov. 1825)

You have all professional deformation of your minds; *convergence* does not matter here ...
(P. Henrici 1985)

We now enter a new era for our subject, more theoretical than the preceding one. It was inaugurated by the work of Cauchy, who was not as fascinated by long numerical calculations as was, say, Euler, but merely a fanatic for perfect mathematical rigor and exactness. He criticized in the work of his predecessors the use of infinite series and other infinite processes without taking much account of error estimates or convergence results. He therefore established around 1820 a convergence theorem for the polygon method of Euler and, some 15 years later, for the power series method of Newton (see Section I.8). Beyond the estimation of errors, these results also allow the statement of *general existence theorems* for the solutions of arbitrary differential equations (“d’une équation différentielle quelconque”), whose solutions were only known before in a very few cases. A second important consequence is to provide results about the *uniqueness* of the solution, which allow one to conclude that the computed solution (numerically or analytically) is the only one with the same initial value and that there are no others. Only then we are allowed to speak of *the* solution of the problem.

His very first proof has recently been discovered on fragmentary notes (Cauchy 1824), which were never published in Cauchy’s lifetime (did his notes not satisfy the Minister of education?: “. . . mais que le second professeur, M. Cauchy, n’a présenté que des feuilles qui n’ont pu satisfaire la commission, et qu’il a été jusqu’à présent impossible de l’amener à se rendre au voeu du Conseil et à exécuter la décision du Ministre”).

Convergence of Euler’s Method

Let us now, with bared head and trembling knees, follow the ideas of this historical proof. We formulate it in a way which generalizes directly to higher dimensional systems.

Starting with the one-dimensional differential equation

$$y' = f(x, y), \quad y(x_0) = y_0, \quad y(X) = ? \quad (7.1)$$

we make use of the method explained by Euler (1768) in the last section of his “Institutiones Calculi Integralis I” (Caput VII, p. 424), i.e., we consider a subdivision

of the interval of integration

$$x_0, x_1, \dots, x_{n-1}, x_n = X \quad (7.2)$$

and replace in each subinterval the solution by the first term of its Taylor series

$$\begin{aligned} y_1 - y_0 &= (x_1 - x_0)f(x_0, y_0) \\ y_2 - y_1 &= (x_2 - x_1)f(x_1, y_1) \\ &\dots \\ y_n - y_{n-1} &= (x_n - x_{n-1})f(x_{n-1}, y_{n-1}). \end{aligned} \quad (7.3)$$

For the subdivision above we also use the notation

$$h = (h_0, h_1, \dots, h_{n-1})$$

where $h_i = x_{i+1} - x_i$. If we connect y_0 and y_1 , y_1 and y_2 , \dots etc by straight lines we obtain the *Euler polygon*

$$y_h(x) = y_i + (x - x_i)f(x_i, y_i) \quad \text{for } x_i \leq x \leq x_{i+1}. \quad (7.3a)$$

Lemma 7.1. Assume that $|f|$ is bounded by A on

$$D = \left\{ (x, y) \mid x_0 \leq x \leq X, |y - y_0| \leq b \right\}.$$

If $X - x_0 \leq b/A$ then the numerical solution (x_i, y_i) given by (7.3), remains in D for every subdivision (7.2) and we have

$$|y_h(x) - y_0| \leq A \cdot |x - x_0|, \quad (7.4)$$

$$\left| y_h(x) - \left(y_0 + (x - x_0)f(x_0, y_0) \right) \right| \leq \varepsilon \cdot |x - x_0| \quad (7.5)$$

if $|f(x, y) - f(x_0, y_0)| \leq \varepsilon$ on D .

Proof. Both inequalities are obtained by adding up the lines of (7.3) and using the triangle inequality. Formula (7.4) then shows immediately that for $A(x - x_0) \leq b$ the polygon remains in D . \square

Our next problem is to obtain an estimate for the change of $y_h(x)$, when the initial value y_0 is changed: let z_0 be another initial value and compute

$$z_1 - z_0 = (x_1 - x_0)f(x_0, z_0). \quad (7.6)$$

We need an estimate for $|z_1 - y_1|$. Subtracting (7.6) from the first line of (7.3) we obtain

$$z_1 - y_1 = z_0 - y_0 + (x_1 - x_0) \left(f(x_0, z_0) - f(x_0, y_0) \right).$$

This shows that we need an estimate for $f(x_0, z_0) - f(x_0, y_0)$. If we suppose

$$|f(x, z) - f(x, y)| \leq L|z - y| \quad (7.7)$$

we obtain

$$|z_1 - y_1| \leq (1 + (x_1 - x_0)L)|z_0 - y_0|. \quad (7.8)$$

Lemma 7.2. *For a fixed subdivision h let $y_h(x)$ and $z_h(x)$ be the Euler polygons corresponding to the initial values y_0 and z_0 , respectively. If*

$$\left| \frac{\partial f}{\partial y}(x, y) \right| \leq L \quad (7.9)$$

in a convex region which contains $(x, y_h(x))$ and $(x, z_h(x))$ for all $x_0 \leq x \leq X$, then

$$|z_h(x) - y_h(x)| \leq e^{L(x-x_0)}|z_0 - y_0|. \quad (7.10)$$

Proof. (7.9) implies (7.7), (7.7) implies (7.8), (7.8) implies

$$|z_1 - y_1| \leq e^{L(x_1-x_0)}|z_0 - y_0|.$$

If we repeat the same argument for $z_2 - y_2$, $z_3 - y_3$, and so on, we finally obtain (7.10). \square

Remark. Condition (7.7) is called a “Lipschitz condition”. It was Lipschitz (1876) who rediscovered the theory (footnote in the paper of Lipschitz: “L’auteur ne connaît pas évidemment les travaux de Cauchy . . .”) and advocated the use of (7.7) instead of the more stringent hypothesis (7.9). Lipschitz’s proof is also explained in the classical work of Picard (1891-96), Vol. II, Chap. XI, Sec. I.

If the subdivision (7.2) is refined more and more, so that

$$|h| := \max_{i=0, \dots, n-1} h_i \rightarrow 0,$$

we expect that the Euler polygons converge to a solution of (7.1). Indeed, we have

Theorem 7.3. *Let $f(x, y)$ be continuous, and $|f|$ be bounded by A and satisfy the Lipschitz condition (7.7) on*

$$D = \{(x, y) \mid x_0 \leq x \leq X, |y - y_0| \leq b\}.$$

If $X - x_0 \leq b/A$, then we have:

- For $|h| \rightarrow 0$ the Euler polygons $y_h(x)$ converge uniformly to a continuous function $\varphi(x)$.*
- $\varphi(x)$ is continuously differentiable and solution of (7.1) on $x_0 \leq x \leq X$.*
- There exists no other solution of (7.1) on $x_0 \leq x \leq X$.*

Proof. a) Take an $\varepsilon > 0$. Since f is uniformly continuous on the compact set D , there exists a $\delta > 0$ such that

$$|u_1 - u_2| \leq \delta \quad \text{and} \quad |v_1 - v_2| \leq A \cdot \delta$$

imply

$$|f(u_1, v_1) - f(u_2, v_2)| \leq \varepsilon. \quad (7.11)$$

Suppose now that the subdivision (7.2) satisfies

$$|x_{i+1} - x_i| \leq \delta, \quad \text{i.e.,} \quad |h| \leq \delta. \quad (7.12)$$

We first study the effect of adding new mesh-points. In a first step, we consider a subdivision $h(1)$, which is obtained by adding new points only to the *first* subinterval (see Fig. 7.1). It follows from (7.5) (applied to this first subinterval) that for the new refined solution $y_{h(1)}(x_1)$ we have the estimate $|y_{h(1)}(x_1) - y_h(x_1)| \leq \varepsilon|x_1 - x_0|$. Since the subdivisions h and $h(1)$ are identical on $x_1 \leq x \leq X$ we can apply Lemma 7.2 to obtain

$$|y_{h(1)}(x) - y_h(x)| \leq e^{L(x-x_1)}(x_1 - x_0)\varepsilon \quad \text{for} \quad x_1 \leq x \leq X.$$

We next add further points to the subinterval (x_1, x_2) and denote the new subdivision by $h(2)$. In the same way as above this leads to $|y_{h(2)}(x_2) - y_{h(1)}(x_2)| \leq \varepsilon|x_2 - x_1|$ and

$$|y_{h(2)}(x) - y_{h(1)}(x)| \leq e^{L(x-x_2)}(x_2 - x_1)\varepsilon \quad \text{for} \quad x_2 \leq x \leq X.$$

The entire situation is sketched in Fig. 7.1. If we denote by \hat{h} the final refinement, we obtain for $x_i < x \leq x_{i+1}$

$$\begin{aligned} |y_{\hat{h}}(x) - y_h(x)| & \leq \varepsilon \left(e^{L(x-x_1)}(x_1 - x_0) + \dots + e^{L(x-x_i)}(x_i - x_{i-1}) \right) + \varepsilon(x - x_i) \\ & \leq \varepsilon \int_{x_0}^x e^{L(x-s)} ds = \frac{\varepsilon}{L} \left(e^{L(x-x_0)} - 1 \right). \end{aligned} \quad (7.13)$$

If we now have two different subdivisions h and \tilde{h} , which both satisfy (7.12), we introduce a *third* subdivision \hat{h} which is a refinement of both subdivisions (just as is usually done in proving the existence of Riemann's integral), and apply (7.13) twice. We then obtain from (7.13) by the triangle inequality

$$|y_h(x) - y_{\tilde{h}}(x)| \leq 2 \frac{\varepsilon}{L} \left(e^{L(x-x_0)} - 1 \right).$$

For $\varepsilon > 0$ small enough, this becomes arbitrarily small and shows the uniform convergence of the Euler polygons to a continuous function $\varphi(x)$.

b) Let

$$\varepsilon(\delta) := \sup \left\{ |f(u_1, v_1) - f(u_2, v_2)| ; |u_1 - u_2| \leq \delta, |v_1 - v_2| \leq A\delta, (u_i, v_i) \in D \right\}$$

be the modulus of continuity. If x belongs to the subdivision h then we obtain from (7.5) (replace (x_0, y_0) by $(x, y_h(x))$ and x by $x + \delta$)

$$|y_h(x + \delta) - y_h(x) - \delta f(x, y_h(x))| \leq \varepsilon(\delta)\delta. \quad (7.14)$$

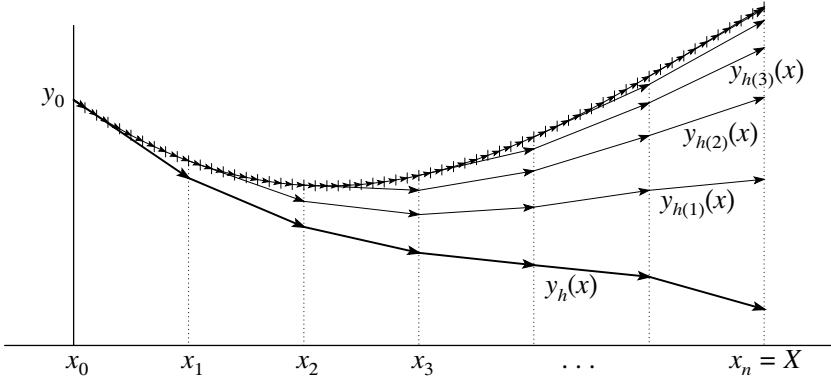


Fig. 7.1. Lady Windermere's Fan (O. Wilde 1892)

Taking the limit $|h| \rightarrow 0$ we get

$$|\varphi(x + \delta) - \varphi(x) - \delta f(x, \varphi(x))| \leq \varepsilon(\delta)\delta. \quad (7.15)$$

Since $\varepsilon(\delta) \rightarrow 0$ for $\delta \rightarrow 0$, this proves the differentiability of $\varphi(x)$ and $\varphi'(x) = f(x, \varphi(x))$.

c) Let $\psi(x)$ be a second solution of (7.1) and suppose that the subdivision h satisfies (7.12). We then denote by $y_h^{(i)}(x)$ the Euler polygon to the initial value $(x_i, \psi(x_i))$ (it is defined for $x_i \leq x \leq X$). It follows from

$$\psi(x) = \psi(x_i) + \int_{x_i}^x f(s, \psi(s)) ds$$

and (7.11) that

$$|\psi(x) - y_h^{(i)}(x)| \leq \varepsilon |x - x_i| \quad \text{for } x_i \leq x \leq x_{i+1}.$$

Using Lemma 7.2 we deduce in the same way as in part a) that

$$|\psi(x) - y_h(x)| \leq \frac{\varepsilon}{L} \left(e^{L(x-x_0)} - 1 \right). \quad (7.16)$$

Taking the limits $|h| \rightarrow 0$ and $\varepsilon \rightarrow 0$ we obtain $|\psi(x) - \varphi(x)| \leq 0$, proving uniqueness. \square

Theorem 7.3 is a *local* existence - and uniqueness - result. However, if we interpret the endpoint of the solution as a new initial value, we can apply Theorem 7.3 again and continue the solution. Repeating this procedure we obtain

Theorem 7.4. Assume U to be an open set in \mathbb{R}^2 and let f and $\partial f / \partial y$ be continuous on U . Then, for every $(x_0, y_0) \in U$, there exists a unique solution of (7.1), which can be continued up to the boundary of U (in both directions).

Proof. Clearly, Theorem 7.3 can be rewritten to give a local existence - and uniqueness - result for an interval (X, x_0) to the left of x_0 . The rest follows from the fact that every point in U has a neighbourhood which satisfies the assumptions of Theorem 7.3. \square

It is interesting to mention that formula (7.13) for $|\widehat{h}| \rightarrow 0$ gives the following *error estimate*

$$|y(x) - y_h(x)| \leq \frac{\varepsilon}{L} \left(e^{L(x-x_0)} - 1 \right) \quad (7.17)$$

for the Euler polygon ($|h| \leq \delta$). Here $y(x)$ stands for the exact solution of (7.1). The next theorem refines the above estimates for the case that $f(x, y)$ is also differentiable with respect to x .

Theorem 7.5. *Suppose that in a neighbourhood of the solution*

$$|f| \leq A, \quad \left| \frac{\partial f}{\partial y} \right| \leq L, \quad \left| \frac{\partial f}{\partial x} \right| \leq M.$$

We then have the following error estimate for the Euler polygons:

$$|y(x) - y_h(x)| \leq \frac{M + AL}{L} \left(e^{L(x-x_0)} - 1 \right) \cdot |h|, \quad (7.18)$$

provided that $|h|$ is sufficiently small.

Proof. For $|u_1 - u_2| \leq |h|$ and $|v_1 - v_2| \leq A|h|$ we obtain, due to the differentiability of f , the estimate

$$|f(u_1, v_1) - f(u_2, v_2)| \leq (M + AL)|h|$$

instead of (7.11). When we insert this amount for ε into (7.16), we obtain the stated result. \square

The estimate (7.18) shows that the global error of Euler's method is proportional to the maximal step size $|h|$. Thus, for an accuracy of, say, three decimal digits, we would need about a thousand steps; a precision of six digits will normally require a million steps etc. We see thus that the present method is not recommended for computations of high precision. In fact, the main subject of Chapter II will be to find methods which converge faster.

Existence Theorem of Peano

Si a est un complexe d'ordre n , et b un nombre réel, alors on peut déterminer b' et f , où b' est une quantité plus grande que b , et f est un signe de fonction qui à chaque nombre de l'intervalle de b à b' fait correspondre un complexe (en d'autres mots, ft est un complexe fonction de la variable réelle t , définie pour toutes les valeurs de l'intervalle (b, b')); la valeur de ft pour $t = b$ est a ; et dans tout l'intervalle (b, b') cette fonction ft satisfait à l'équation différentielle donnée. (Original version of Peano's Theorem)

The Lipschitz condition (7.7) is a crucial tool in the proof of (7.10) and finally of the Convergence Theorem. If we completely abandon condition (7.7) and only require that $f(x, y)$ be continuous, the convergence of the Euler polygons is no longer guaranteed.

An example, plotted in Fig. 7.2, is given by the equation

$$y' = 4 \left(\text{sign}(y) \sqrt{|y|} + \max\left(0, x - \frac{|y|}{x}\right) \cdot \cos\left(\frac{\pi \log x}{\log 2}\right) \right) \quad (7.19)$$

with $y(0) = 0$. It has been constructed such that

$$\begin{aligned} f(h, 0) &= 4(-1)^i h & \text{for } h = 2^{-i}, \\ f(x, y) &= 4 \text{sign}(y) \cdot \sqrt{|y|} & \text{for } |y| \geq x^2. \end{aligned}$$

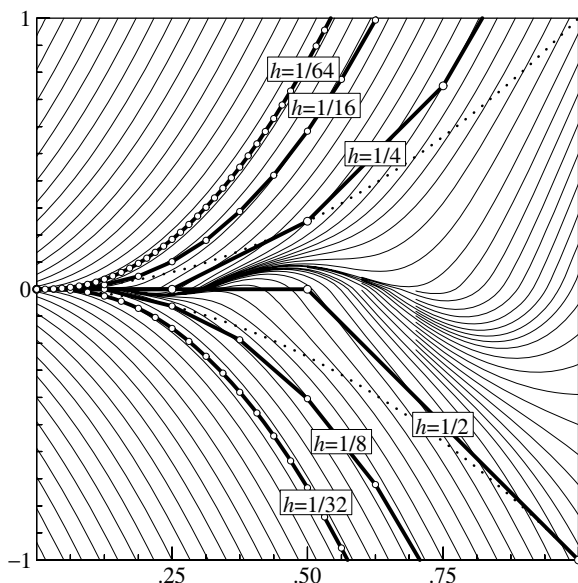


Fig. 7.2. Solution curves and Euler polygons for equation (7.19)

There is an infinity of solutions for this initial value, some of which are plotted in Fig. 7.2. The Euler polygons converge for $h = 2^{-i}$ and even i to the maximal solution $y = 4x^2$, and for odd i to $y = -4x^2$. For other sequences of h all intermediate solutions can be obtained as well.

Theorem 7.6 (Peano 1890). *Let $f(x, y)$ be continuous and $|f|$ be bounded by A on*

$$D = \{(x, y) \mid x_0 \leq x \leq X, |y - y_0| \leq b\}.$$

If $X - x_0 \leq b/A$, then there is a subsequence of the sequence of the Euler polygons which converges to a solution of the differential equation.

The original proof of Peano is, in its crucial part on the convergence result, very brief and not clear to unexperienced readers such as us. Arzelà (1895), who took up the subject again, explains his ideas in more detail and emphasizes the need for an *equicontinuity* of the sequence. The proof usually given nowadays (for what has become the theorem of Arzelà-Ascoli), was only introduced later (see e.g. Perron (1918), Hahn (1921), p. 303) and is sketched as follows:

Proof. Let

$$v_1(x), v_2(x), v_3(x), \dots \quad (7.20)$$

be a sequence of Euler polygons for decreasing step sizes. It follows from (7.4) that for fixed x this sequence is bounded. We choose a sequence of numbers r_1, r_2, r_3, \dots dense in the interval (x_0, X) . There is now a subsequence of (7.20) which converges for $x = r_1$ (Bolzano-Weierstrass), say

$$v_1^{(1)}(x), v_2^{(1)}(x), v_3^{(1)}(x), \dots \quad (7.21)$$

We next select a subsequence of (7.21) which converges for $x = r_2$

$$v_1^{(2)}(x), v_2^{(2)}(x), v_3^{(2)}(x), \dots \quad (7.22)$$

and so on. Then take the “diagonal” sequence

$$v_1^{(1)}(x), v_2^{(2)}(x), v_3^{(3)}(x), \dots \quad (7.23)$$

which, apart from a finite number of terms, is a subsequence of each of these sequences, and thus converges for all r_i . Finally, with the estimate

$$|v_n^{(n)}(x) - v_n^{(n)}(r_j)| \leq A|x - r_j|$$

(see (7.4)), which expresses the equicontinuity of the sequence, we obtain

$$\begin{aligned} & |v_n^{(n)}(x) - v_m^{(m)}(x)| \\ & \leq |v_n^{(n)}(x) - v_n^{(n)}(r_j)| + |v_n^{(n)}(r_j) - v_m^{(m)}(r_j)| + |v_m^{(m)}(r_j) - v_m^{(m)}(x)| \\ & \leq 2A|x - r_j| + |v_n^{(n)}(r_j) - v_m^{(m)}(r_j)|. \end{aligned}$$

For fixed $\varepsilon > 0$ we then choose a finite subset R of $\{r_1, r_2, \dots\}$ satisfying

$$\min\{|x - r_j|; r_j \in R, x_0 \leq x \leq X\} \leq \varepsilon/A$$

and secondly we choose N such that

$$|v_n^{(n)}(r_j) - v_m^{(m)}(r_j)| \leq \varepsilon \quad \text{for } n, m \geq N \quad \text{and } r_j \in R.$$

This shows the uniform convergence of (7.23). In the same way as in part b) of the proof of Theorem 7.3 it follows that the limit function is a solution of (7.1). One only has to add an $\mathcal{O}(|h|)$ -term in (7.14), if x is not a subdivision point. \square

Exercises

1. Apply Euler's method with constant step size $x_{i+1} - x_i = 1/n$ to the differential equation $y' = ky$, $y(0) = 1$ and obtain a classical approximation for the solution $y(1) = e^k$. Give an estimate of the error.
2. Apply Euler's method with constant step size to
 - a) $y' = y^2$, $y(0) = 1$, $y(1/2) = ?$
 - b) $y' = x^2 + y^2$, $y(0) = 0$, $y(1/2) = ?$

Make rigorous error estimates using Theorem 7.4 and compare these estimates with the actual errors. The main difficulty is to find a suitable region in which the estimates of Theorem 7.4 hold, without making the constants A , L , M too large and, at the same time, ensuring that the solution curves remain inside this region (see also I.8, Exercise 3).

3. Prove the result: if the differential equation $y' = f(x, y)$, $y(x_0) = y_0$ with f continuous, possesses a unique solution, then the Euler polygons converge to this solution.
4. "There is an elementary proof of Peano's existence theorem" (Walter 1971). Suppose that A is a bound for $|f|$. Then the sequence

$$y_{i+1} = y_i + h \cdot \max\{f(x, y) | x_i \leq x \leq x_{i+1}, y_i - 3Ah \leq y \leq y_i + Ah\}$$

converges for all continuous f to a (the maximal) solution. Try to prove this. Unfortunately, this proof does not extend to systems of equations, unless they are "quasimonotone" (see Section I.10, Exercise 3).

I.8 Existence Theory using Iteration Methods and Taylor Series

A second approach to existence theory is possible with the help of an iterative refinement of approximate solutions. The first appearances of the idea are very old. For instance many examples of this type can be found in the work of Lagrange, above all in his astronomical calculations. Let us consider here the following illustrative example of a Riccati equation

$$y' = x^2 + y + 0.1y^2, \quad y(0) = 0. \quad (8.1)$$

Because of the quadratic term, there is no elementary solution. A very natural idea is therefore to neglect this term, which is in fact very small at the beginning, and to solve for the moment

$$y_1' = x^2 + y_1, \quad y_1(0) = 0. \quad (8.2)$$

This gives, with formula (3.3), a first approximation

$$y_1(x) = 2e^x - (x^2 + 2x + 2). \quad (8.3)$$

With the help of this solution, we now know more about the initially neglected term $0.1y^2$; it will be close to $0.1y_1^2$. So the idea lies at hand to reintroduce this solution into (8.1) and solve now the differential equation

$$y_2' = x^2 + y_2 + 0.1 \cdot (y_1(x))^2, \quad y_2(0) = 0. \quad (8.4)$$

We can use formula (3.3) again and obtain after some calculations

$$y_2(x) = y_1(x) + \frac{2}{5}e^{2x} - \frac{2}{15}e^x(x^3 + 3x^2 + 6x - 54) - \frac{1}{10}(x^4 + 8x^3 + 32x^2 + 72x + 76).$$

This is already much closer to the correct solution, as can be seen from the following comparison of the errors $e_1 = y(x) - y_1(x)$ and $e_2 = y(x) - y_2(x)$:

$x = 0.2$	$e_1 = 0.228 \times 10^{-07}$	$e_2 = 0.233 \times 10^{-12}$
$x = 0.4$	$e_1 = 0.327 \times 10^{-05}$	$e_2 = 0.566 \times 10^{-09}$
$x = 0.8$	$e_1 = 0.534 \times 10^{-03}$	$e_2 = 0.165 \times 10^{-05}$.

It looks promising to continue this process, but the computations soon become very tedious.

Picard-Lindelöf Iteration

The general formulation of the method is the following: we try, if possible, to split up the function $f(x, y)$ of the differential equation

$$y' = f(x, y) = f_1(x, y) + f_2(x, y), \quad y(x_0) = y_0 \quad (8.5)$$

so that any differential equation of the form $y' = f_1(x, y) + g(x)$ can be solved analytically and so that $f_2(x, y)$ is small. Then we start with a first approximation $y_0(x)$ and compute successively $y_1(x), y_2(x), \dots$ by solving

$$y'_{i+1} = f_1(x, y_{i+1}) + f_2(x, y_i(x)), \quad y_{i+1}(x_0) = y_0. \quad (8.6)$$

The most primitive form of this process is obtained by choosing $f_1 = 0$, $f_2 = f$, in which case (8.6) is immediately integrated and becomes

$$y_{i+1}(x) = y_0 + \int_{x_0}^x f(s, y_i(s)) ds. \quad (8.7)$$

This is called the *Picard-Lindelöf iteration method*. It appeared several times in the literature, e.g., in Liouville (1838), Cauchy, Peano (1888), Lindelöf (1894), Bendixson (1893). Picard (1890) considered it merely as a by-product of a similar idea for partial differential equations and analyzed it thoroughly in his famous treatise Picard (1891-96), Vol. II, Chap. XI, Sect. III.

The fast *convergence* of the method, for $|x - x_0|$ small, is readily seen: if we subtract formula (8.7) from the same with i replaced by $i - 1$, we have

$$y_{i+1}(x) - y_i(x) = \int_{x_0}^x \left(f(s, y_i(s)) - f(s, y_{i-1}(s)) \right) ds. \quad (8.8)$$

We now apply the Lipschitz condition (7.7) and the triangle inequality to obtain

$$|y_{i+1}(x) - y_i(x)| \leq L \int_{x_0}^x |y_i(s) - y_{i-1}(s)| ds. \quad (8.9)$$

When we assume $y_0(x) \equiv y_0$, the triangle inequality applied to (8.7) with $i = 0$ yields the estimate

$$|y_1(x) - y_0(x)| \leq A|x - x_0|$$

where A is a bound for $|f|$ as in Section I.7. We next insert this into the right hand side of (8.9) repeatedly to obtain finally the estimate (Lindelöf 1894)

$$|y_i(x) - y_{i-1}(x)| \leq AL^{i-1} \frac{|x - x_0|^i}{i!}. \quad (8.10)$$

The right-hand side is a term of the Taylor series for $e^{L|x-x_0|}$, which converges for all x ; we therefore conclude that $|y_{i+k} - y_i|$ becomes arbitrarily small when i is large. The error is bounded by the remainder of the above exponential series. So the sequence $y_i(x)$ converges uniformly to the solution $y(x)$. For example, if $L|x - x_0| \leq 1/10$ and the constant A is moderate, 10 iterations would provide a numerical solution with about 17 correct digits.

The main practical drawback of the method is the need for repeated computation of integrals, which is usually not very convenient, if at all analytically possible, and soon becomes very tedious. However, its fast convergence and new machine architectures (parallelism) coupled with numerical evaluations of the integrals have made the approach interesting for large problems (see Nevanlinna 1989).

Taylor Series

Après avoir montré l'insuffisance des méthodes d'intégration fondées sur le développement en séries, il me reste à dire en peu de mots ce qu'on peut leur substituer. (Cauchy)

A third existence proof can be based on a study of the convergence of the Taylor series of the solutions. This was mentioned in a footnote of Liouville (1836, p. 255), and brought to perfection by Cauchy (1839-42).

We have already seen the recursive computation of the Taylor coefficients in the work of Newton (see Section I.2). Euler (1768) then formulated the general procedure for the higher derivatives of the solution of

$$y' = f(x, y), \quad y(x_0) = y_0 \quad (8.11)$$

which, by successive differentiation, are obtained as

$$\begin{aligned} y'' &= f_x + f_y y' = f_x + f_y f \\ y''' &= f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_y(f_x + f_y f) \end{aligned} \quad (8.12)$$

etc. Then the solution is

$$y(x_0 + h) = y(x_0) + y'(x_0)h + y''(x_0)\frac{h^2}{2!} + \dots \quad (8.13)$$

The formulas (8.12) for higher derivatives soon become very complicated. Euler therefore proposed to use only a few terms of this series with h sufficiently small and to repeat the computations from the point $x_1 = x_0 + h$ ("analytic continuation").

We shall now outline the main ideas of Cauchy's *convergence proof* for the series (8.13). We suppose that $f(x, y)$ is *analytic* in the neighbourhood of the initial value x_0, y_0 , which for simplicity of notation we assume located at the origin $x_0 = y_0 = 0$:

$$f(x, y) = \sum_{i,j \geq 0} a_{ij} x^i y^j, \quad (8.14)$$

where the a_{ij} are multiples of the partial derivatives occurring in (8.12). If the series (8.14) is assumed to converge for $|x| \leq r$, $|y| \leq r$, then the Cauchy inequalities from classical complex analysis give

$$|a_{ij}| \leq \frac{M}{r^{i+j}}, \quad \text{where} \quad M = \max_{|x| \leq r, |y| \leq r} |f(x, y)|. \quad (8.15)$$

The idea is now the following: since all signs in (8.12) are positive, we obtain the worst possible result if we replace in (8.14) all a_{ij} by the largest possible values (8.15) (“method of majorants”):

$$f(x, y) \rightarrow \sum_{i,j \geq 0} M \frac{x^i y^j}{r^{i+j}} = \frac{M}{(1-x/r)(1-y/r)}.$$

However, the majorizing differential equation

$$y' = \frac{M}{(1-x/r)(1-y/r)}, \quad y(0) = 0$$

is readily integrated by separation of variables (see Section I.3) and has the solution

$$y = r \left(1 - \sqrt{1 + 2M \log \left(1 - \frac{x}{r} \right)} \right). \quad (8.16)$$

This solution has a power series expansion which converges for all x such that $|2M \log(1 - x/r)| < 1$. Therefore, the series (8.13) also converges at least for all $|h| < r(1 - \exp(-1/2M))$. \square

Recursive Computation of Taylor Coefficients

... dieses Verfahren praktisch nicht in Frage kommen kann.
(Runge & König 1924)

The exact opposite is true, if we use the right approach ...
(R.E. Moore 1979)

The “right approach” is, in fact, an extension of Newton’s approach and has been rediscovered several times (e.g., Steffensen 1956) and implemented into computer programs by Gibbons (1960) and Moore (1966). For a more extensive bibliography see the references in Wanner (1969), p. 10-20.

The idea is the following: let

$$Y_i = \frac{1}{i!} y^{(i)}(x_0), \quad F_i = \frac{1}{i!} \left(f(x, y(x)) \right)^{(i)} \Big|_{x=x_0} \quad (8.17)$$

be the Taylor coefficients of $y(x)$ and of $f(x, y(x))$, so that (8.13) becomes

$$y(x_0 + h) = \sum_{i=0}^{\infty} h^i Y_i.$$

Then, from (8.11),

$$Y_{i+1} = \frac{1}{i+1} F_i. \quad (8.18)$$

Now suppose that $f(x, y)$ is the composition of a sequence of algebraic operations and elementary functions. This leads to a sequence of items,

$$x, y, p, q, r, \dots, \text{ and finally } f. \quad (8.19)$$

For each of these items we find formulas for generating the i th Taylor coefficient from the preceding ones as follows:

a) $r = p \pm q$:

$$R_i = P_i \pm Q_i, \quad i = 0, 1, \dots \quad (8.20a)$$

b) $r = pq$: the Cauchy product yields

$$R_i = \sum_{j=0}^i P_j Q_{i-j}, \quad i = 0, 1, \dots \quad (8.20b)$$

c) $r = p/q$: write $p = rq$, use formula b) and solve for R_i :

$$R_i = \frac{1}{Q_0} \left(P_i - \sum_{j=0}^{i-1} R_j Q_{i-j} \right), \quad i = 0, 1, \dots \quad (8.20c)$$

There also exist formulas for many elementary functions (in fact, because these functions are themselves solutions of rational differential equations).

d) $r = \exp(p)$: use $r' = p' \cdot r$ and apply (8.20b). This gives for $i = 1, 2, \dots$

$$R_0 = \exp(P_0), \quad R_i = \frac{1}{i} \sum_{j=0}^{i-1} (i-j) R_j P_{i-j}. \quad (8.20d)$$

e) $r = \log(p)$: use $p = \exp(r)$ and rearrange formula d). This gives

$$R_0 = \log(P_0), \quad R_i = \frac{1}{P_0} \left(P_i - \frac{1}{i} \sum_{j=1}^{i-1} (i-j) P_j R_{i-j} \right). \quad (8.20e)$$

f) $r = p^c$, $c \neq 1$ constant. Use $pr' = crp'$ and apply (8.20b):

$$R_0 = P_0^c, \quad R_i = \frac{1}{iP_0} \left(\sum_{j=0}^{i-1} (ci - (c+1)j) R_j P_{i-j} \right). \quad (8.20f)$$

g) $r = \cos(p)$, $s = \sin(p)$: as in d) we have

$$\begin{aligned} R_0 &= \cos P_0, & R_i &= -\frac{1}{i} \sum_{j=0}^{i-1} (i-j) S_j P_{i-j}, \\ S_0 &= \sin P_0, & S_i &= \frac{1}{i} \sum_{j=0}^{i-1} (i-j) R_j P_{i-j}. \end{aligned} \quad (8.20g)$$

The alternating use of (8.20) and (8.18) then allows us to compute the Taylor coefficients for (8.17) to any wanted order in a very economical way. It is not difficult to write subroutines for the above formulas, which have to be called in the same order as the differential equation (8.11) is composed of elementary operations. There also exist computer programs which “compile” Fortran statements for $f(x, y)$ into this list of subroutine calls. One has been written by T. Szymanski and J.H. Gray (see Knapp & Wanner 1969).

Example. The differential equation $y' = x^2 + y^2$ leads to the recursion

$$Y_0 = y(0), \quad Y_{i+1} = \frac{1}{i+1} \left(P_i + \sum_{j=0}^i Y_j Y_{i-j} \right), \quad i = 0, 1, \dots$$

where $P_i = 1$ for $i = 2$ and $P_i = 0$ for $i \neq 2$ are the coefficients for x^2 . One can imagine how much easier this is than formulas (8.12).

An important property of this approach is that it can be executed in *interval analysis* and thus allows us to obtain *reliable error bounds* by the use of Lagrange’s error formula for Taylor series. We refer to the books by R.E. Moore (1966) and (1979) for more details.

Exercises

1. Obtain from (8.10) the estimate

$$|y_i(x) - y_0| \leq \frac{A}{L} \left(e^{L(x-x_0)} - 1 \right)$$

and explain the similarity of this result with (7.16).

2. Apply the method of Picard to the problem $y' = Ky$, $y(0) = 1$.
3. Compute three Picard iterations for the problem $y' = x^2 + y^2$, $y(0) = 0$, $y(1/2) = ?$ and make a rigorous error estimate. Compare the result with the correct solution $y(1/2) = 0.041791146154681863220768806849179$.

4. Compute with an iteration method the solution of

$$y' = \sqrt{x} + \sqrt{y}, \quad y(0) = 0$$

and observe that the method can work well for equations which pose serious problems with other methods. An even greater difference occurs for the equations

$$y' = \sqrt{x} + y^2, \quad y(0) = 0 \quad \text{and} \quad y' = \frac{1}{\sqrt{x}} + y^2, \quad y(0) = 0.$$

5. Define $f(x, y)$ by

$$f(x, y) = \begin{cases} 0 & \text{for } x \leq 0 \\ 2x & \text{for } x > 0, y < 0 \\ 2x - \frac{4y}{x} & \text{for } 0 \leq y \leq x^2 \\ -2x & \text{for } x > 0, x^2 < y. \end{cases}$$

- a) Show that $f(x, y)$ is continuous, but not Lipschitz.
 - b) Show that for the problem $y' = f(x, y)$, $y(0) = 0$ the Picard iteration method does not converge.
 - c) Show that there is a unique solution and that the Euler polygons converge.
6. Use the method of Picard iteration to prove: if $f(x, y)$ is continuous and satisfies a Lipschitz condition (7.7) on the infinite strip $D = \{(x, y) ; x_0 \leq x \leq X\}$, then the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ possesses a unique solution on $x_0 \leq x \leq X$.

Compare this global result with Theorem 7.3.

7. Define a function $y(x)$ (the “inverse error function”) by the relation

$$x = \frac{2}{\sqrt{\pi}} \int_0^y e^{-t^2} dt$$

and show that it satisfies the differential equation

$$y' = \frac{\sqrt{\pi}}{2} e^{y^2}, \quad y(0) = 0.$$

Obtain recursion formulas for its Taylor coefficients.

I.9 Existence Theory for Systems of Equations

The first treatment of an existence theory for simultaneous systems of differential equations was undertaken in the last existing pages (p. 123-136) of Cauchy (1824). We write the equations as

$$\begin{aligned} y_1' &= f_1(x, y_1, \dots, y_n), & y_1(x_0) &= y_{10}, & y_1(X) &=? \\ \dots & & \dots & & \dots & \\ y_n' &= f_n(x, y_1, \dots, y_n), & y_n(x_0) &= y_{n0}, & y_n(X) &=? \end{aligned} \quad (9.1)$$

and ask for the existence of the n solutions $y_1(x), \dots, y_n(x)$. It is again natural to consider, in analogy to (7.3), the method of Euler

$$y_{k,i+1} = y_{ki} + (x_{i+1} - x_i) \cdot f_k(x_i, y_{1i}, \dots, y_{ni}) \quad (9.2)$$

(for $k = 1, \dots, n$ and $i = 0, 1, 2, \dots$). Here y_{ki} is intended to approximate $y_k(x_i)$, where $x_0 < x_1 < x_2 \dots$ is a subdivision of the interval of integration as in (7.2).

We now try to carry over everything we have done in Section I.7 to the new situation. Although we have no problem in extending (7.4) to the estimate

$$|y_{ki} - y_{k0}| \leq A_k |x_i - x_0| \quad \text{if} \quad |f_k(x, y_1, \dots, y_n)| \leq A_k, \quad (9.3)$$

things become a little more complicated for (7.7): we have to estimate

$$f_k(x, z_1, \dots, z_n) - f_k(x, y_1, \dots, y_n) = \frac{\partial f_k}{\partial y_1} \cdot (z_1 - y_1) + \dots + \frac{\partial f_k}{\partial y_n} \cdot (z_n - y_n), \quad (9.4)$$

where the derivatives $\partial f_k / \partial y_i$ are taken at suitable intermediate points. Here Cauchy uses the inequality now called the “Cauchy-Schwarz inequality” (“Enfin, il résulte de la formule (13) de la 11e leçon du calcul différentiel ...”) to obtain

$$\begin{aligned} & |f_k(x, z_1, \dots, z_n) - f_k(x, y_1, \dots, y_n)| \\ & \leq \sqrt{\left(\frac{\partial f_k}{\partial y_1}\right)^2 + \dots + \left(\frac{\partial f_k}{\partial y_n}\right)^2} \cdot \sqrt{(z_1 - y_1)^2 + \dots + (z_n - y_n)^2}. \end{aligned} \quad (9.5)$$

At this stage, we begin to feel that further development is advisable only after the introduction of vector notation.

Vector Notation

This was promoted in our subject by the papers of Peano, (1888) and (1890), who was influenced, as he says, by the famous “Ausdehnungslehre” of Grassmann and the work of Hamilton, Cayley, and Sylvester. We introduce the vectors (Peano called them “complexes”)

$$y = (y_1, \dots, y_n)^T, \quad y_i = (y_{1i}, \dots, y_{ni})^T, \quad z = (z_1, \dots, z_n)^T \quad \text{etc,}$$

and hope that the reader will not confuse the components y_i of a vector y with vectors with indices. We consider the “vector function”

$$f(x, y) = (f_1(x, y), \dots, f_n(x, y))^T,$$

so that equations (9.1) become

$$y' = f(x, y), \quad y(x_0) = y_0, \quad y(X) = ?, \quad (9.1')$$

Euler’s method (9.2) is

$$y_{i+1} = y_i + (x_{i+1} - x_i)f(x_i, y_i), \quad i = 0, 1, 2, \dots \quad (9.2')$$

and the Euler polygon is given by

$$y_h(x) = y_i + (x - x_i)f(x_i, y_i) \quad \text{for} \quad x_i \leq x \leq x_{i+1}.$$

There is no longer any difference in notation with the one-dimensional cases (7.1), (7.3) and (7.3a).

In view of estimate (9.5), we introduce for a vector $y = (y_1, \dots, y_n)^T$ the *norm* (originally “modulus”)

$$\|y\| = \sqrt{y_1^2 + \dots + y_n^2} \quad (9.6)$$

which satisfies all the usual properties of a norm, for example the triangle inequality

$$\|y + z\| \leq \|y\| + \|z\|, \quad \left\| \sum_{i=1}^n y_i \right\| \leq \sum_{i=1}^n \|y_i\|. \quad (9.7)$$

The Euclidean norm (9.6) is not the only one possible, we also use (“on pourrait aussi définir par *mx* la plus grande des valeurs absolues des éléments de x ; alors les propriétés des modules sont presque évidentes.”, Peano)

$$\|y\| = \max(|y_1|, \dots, |y_n|), \quad (9.6')$$

$$\|y\| = |y_1| + \dots + |y_n|. \quad (9.6'')$$

We are now able to formulate estimate (9.3) as follows, in perfect analogy with (7.4): if for some norm $\|f(x, y)\| \leq A$ on $D = \{(x, y) \mid x_0 \leq x \leq X, \|y - y_0\| \leq b\}$ and if $X - x_0 \leq b/A$ then the numerical solution (x_i, y_i) , given by (9.2'), remains in D and we have

$$\|y_h(x) - y_0\| \leq A \cdot |x - x_0|. \quad (9.8)$$

The analogue of estimate (7.5) can be obtained similarly.

In order to prove the implication “(7.9) \Rightarrow (7.7)” for vector-valued functions it is convenient to work with norms of matrices.

Subordinate Matrix Norms

The relation (9.4) shows that the difference $f(x, z) - f(x, y)$ can be written as the product of a matrix with the vector $z - y$. It is therefore of interest to estimate $\|Qv\|$ and to find the best possible estimate of the form $\|Qv\| \leq \beta\|v\|$.

Definition 9.1. Let Q be a matrix (n columns, m rows) and $\|\dots\|$ be one of the norms defined in (9.6), (9.6') or (9.6''). The *subordinate matrix norm* of Q is then defined by

$$\|Q\| = \sup_{v \neq 0} \frac{\|Qv\|}{\|v\|} = \sup_{\|u\|=1} \|Qu\|. \quad (9.9)$$

By definition, $\|Q\|$ is the smallest number such that

$$\|Qv\| \leq \|Q\| \cdot \|v\| \quad \text{for all } v \quad (9.10)$$

holds. The following theorem gives explicit formulas for the computation of (9.9).

Theorem 9.2. *The norm of a matrix Q is given by the following formulas: for the Euclidean norm (9.6),*

$$\|Q\| = \sqrt{\text{largest eigenvalue of } Q^T Q}; \quad (9.11)$$

for the max-norm (9.6'),

$$\|Q\| = \max_{k=1, \dots, m} \left(\sum_{i=1}^n |q_{ki}| \right); \quad (9.11')$$

for the norm (9.6''),

$$\|Q\| = \max_{i=1, \dots, n} \left(\sum_{k=1}^m |q_{ki}| \right). \quad (9.11'')$$

Proof. Formula (9.11) can be seen from $\|Qv\|^2 = v^T Q^T Q v$ with the help of an orthogonal transformation of $Q^T Q$ to diagonal form.

Formula (9.11') is obtained as follows (we denote (9.6') by $\|\dots\|_\infty$):

$$\|Qv\|_\infty = \max_{k=1, \dots, m} \left| \sum_{i=1}^n q_{ki} v_i \right| \leq \left(\max_{k=1, \dots, m} \sum_{i=1}^n |q_{ki}| \right) \cdot \|v\|_\infty \quad (9.12)$$

shows that $\|Q\| \leq \max_k \sum_i |q_{ki}|$. The equality in (9.11') is then seen by choosing a vector of the form $v = (\pm 1, \pm 1, \dots, \pm 1)^T$ for which equality holds in (9.12). The formula (9.11'') is proved along the same lines. \square

All these formulas remain valid for *complex matrices*. Q^T has only to be replaced by Q^* (transposed and complex conjugate). See e.g., Wilkinson (1965), p. 55-61, Bakhvalov (1976), Chap. VI, Par. 3. With these preparations it is possible to formulate the desired estimate.

Theorem 9.3. *If $f(x, y)$ is differentiable with respect to y in an open convex region U and if*

$$\left\| \frac{\partial f}{\partial y}(x, y) \right\| \leq L \quad \text{for} \quad (x, y) \in U \quad (9.13)$$

then

$$\|f(x, z) - f(x, y)\| \leq L \|z - y\| \quad \text{for} \quad (x, y), (x, z) \in U. \quad (9.14)$$

(Obviously, the matrix norm in (9.13) is subordinate to the norm used in (9.14).)

Proof. This is the “mean value theorem” and its proof can be found in every textbook on calculus. In the case where $\partial f / \partial y$ is continuous, the following simple proof is possible. We consider $\varphi(t) = f(x, y + t(z - y))$ and integrate its derivative (componentwise) from 0 to 1

$$\begin{aligned} f(x, z) - f(x, y) &= \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t) dt \\ &= \int_0^1 \frac{\partial f}{\partial y}(x, y + t(z - y)) \cdot (z - y) dt. \end{aligned} \quad (9.15)$$

Taking the norm of (9.15), using

$$\left\| \int_0^1 g(t) dt \right\| \leq \int_0^1 \|g(t)\| dt, \quad (9.16)$$

and applying (9.10) and (9.13) yields the estimate (9.14). The relation (9.16) is proved by applying the triangle inequality (9.7) to the finite Riemann sums which define the two integrals. \square

We thus have obtained the analogue of (7.7). All that remains to do is, *Da capo al fine*, to read Sections I.7 and I.8 again: *Lemma 7.2, Theorems 7.3, 7.4, 7.5, and 7.6 together with their proofs and the estimates (7.10), (7.13), (7.15), (7.16), (7.17), and (7.18) carry over to the more general case with the only changes that some absolute values are to be replaced by norms.*

The *Picard-Lindelöf iteration* also carries over to systems of equations when in (8.7) we interpret $y_{i+1}(x)$, y_0 and $f(s, y_i(s))$ as vectors, integrated componentwise. The convergence result with the estimate (8.10) also remains the same; for its proof we have to use, between (8.8) and (8.9), the inequality (9.16).

The Taylor series method, its convergence proof, and the recursive generation of the Taylor coefficients also generalize in a straightforward manner to systems of equations.

Exercises

1. Solve the system

$$\begin{aligned} y_1' &= -y_2, & y_1(0) &= 1 \\ y_2' &= +y_1, & y_2(0) &= 0 \end{aligned}$$

by the methods of Euler and Picard, establish rigorous error estimates for all three norms mentioned. Verify the results using the correct solution $y_1(x) = \cos x$, $y_2(x) = \sin x$.

2. Consider the differential equations

$$\begin{aligned} y_1' &= -100y_1 + y_2, & y_1(0) &= 1, & y_1(1) &=? \\ y_2' &= y_1 - 100y_2, & y_2(0) &= 0, & y_2(1) &=? \end{aligned}$$

- Compute the exact solution $y(x)$ by the method explained in Section I.6.
 - Compute the error bound for $\|z(x) - y(x)\|$, where $z(x) = 0$, obtained from (7.10).
 - Apply the method of Euler to this equation with $h = 1/10$.
 - Apply Picard's iteration method.
3. Compute the Taylor series solution of the system with constant coefficients $y' = Ay$, $y(0) = y_0$. Prove that this series converges for all x . Apply this series to the equation of Exercise 1.

Result.

$$y(x) = \sum_{i=0}^{\infty} \frac{x^i}{i!} A^i y_0 =: e^{Ax} y_0.$$

I.10 Differential Inequalities

Differential inequalities are an elegant instrument for gaining a better understanding of equations (7.10), (7.17) and much new insight. This subject was inaugurated in the paper, once again, Peano (1890) and further developed by Perron (1915), Müller (1926), Kamke (1930). A classical treatise on the subject is the book of Walter (1970).

Introduction

The basic idea is the following: let $v(x)$ denote the Euler polygon defined in (7.3) or (9.2), so that

$$v'(x) = f(x_i, y_i) \quad \text{for} \quad x_i < x < x_{i+1}. \quad (10.1)$$

For any chosen norm, we investigate the *error*

$$m(x) = \|v(x) - y(x)\| \quad (10.2)$$

as a function of x and we naturally try to estimate its growth.

Unfortunately, $m(x)$ is not necessarily differentiable, due firstly to the corners of the Euler polygons and secondly, to corners originating from the norms, especially the norms (9.6') and (9.6''). Therefore we consider the so-called *Dini derivatives* defined by

$$D^+m(x) = \limsup_{h \rightarrow 0, h > 0} \frac{m(x+h) - m(x)}{h},$$

$$D_+m(x) = \liminf_{h \rightarrow 0, h > 0} \frac{m(x+h) - m(x)}{h},$$

(see e.g., Scheeffer (1884), Hobson (1921), Chap. V, §260, §280). The property

$$\|w(x+h)\| - \|w(x)\| \leq \|w(x+h) - w(x)\| \quad (10.3)$$

is a simple consequence of the triangle inequality (9.7). If we divide (10.3) by $h > 0$, we obtain the estimates

$$D_+\|w(x)\| \leq \|w'(x+0)\|, \quad D^+\|w(x)\| \leq \|w'(x+0)\|, \quad (10.4)$$

where $w'(x+0)$ is the right derivative of the vector function $w(x)$. If we apply this to $m(x)$ of (10.2), we obtain

$$\begin{aligned} D_+ m(x) &\leq \|v'(x+0) - y'(x)\| \\ &= \|v'(x+0) - f(x, v(x)) + f(x, v(x)) - f(x, y(x))\| \end{aligned}$$

and, using the triangle inequality and the Lipschitz condition (9.14),

$$D_+ m(x) \leq \delta(x) + L \cdot m(x). \quad (10.5)$$

Here, we have introduced

$$\delta(x) = \|v'(x+0) - f(x, v(x))\| \quad (10.6)$$

which is called the *defect* of the approximate solution $v(x)$. This fundamental quantity measures the extent to which the function $v(x)$ does *not* satisfy the imposed differential equation. (7.11) together with (10.1) tell us that $\delta(x) \leq \varepsilon$, so that (10.5) can be further estimated to become

$$D_+ m(x) \leq L \cdot m(x) + \varepsilon, \quad m(x_0) = 0. \quad (10.7)$$

Formula (10.7) (or (10.5)) is what one calls a *differential inequality*. The question is: are we allowed to replace “ \leq ” by “ $=$ ”, i.e., to solve instead of (10.7) the equation

$$u' = Lu + \varepsilon, \quad u(x_0) = 0 \quad (10.8)$$

and to conclude that $m(x) \leq u(x)$? This would mean, by the formulas of Section I.3 or I.5, that

$$m(x) \leq \frac{\varepsilon}{L} \left(e^{L(x-x_0)} - 1 \right). \quad (10.9)$$

We would thus have obtained (7.17) in a natural way and have furthermore discovered an elegant and powerful tool for many kinds of new estimates.

The Fundamental Theorems

A general theorem of the type

$$\left. \begin{aligned} D_+ m(x) &\leq g(x, m(x)) \\ D_+ u(x) &\geq g(x, u(x)) \\ m(x_0) &\leq u(x_0) \end{aligned} \right\} \implies m(x) \leq u(x) \quad \text{for } x_0 \leq x \quad (10.10)$$

cannot be true. Counter-examples are provided by any differential equation with non-unique solutions, such as

$$g(x, y) = \sqrt{y}, \quad m(x) = \frac{x^2}{4}, \quad u(x) = 0. \quad (10.11)$$

The important observation, due to Peano and Perron, which allows us to overcome this difficulty, is that *one* of the first two inequalities must be replaced by a *strict* inequality (see Peano (1890), §3, Lemme 1):

Theorem 10.1. *Suppose that the functions $m(x)$ and $u(x)$ are continuous and satisfy for $x_0 \leq x < X$*

$$\begin{aligned} a) \quad & D_+ m(x) \leq g(x, m(x)) \\ b) \quad & D_+ u(x) > g(x, u(x)) \\ c) \quad & m(x_0) \leq u(x_0). \end{aligned} \tag{10.12}$$

Then

$$m(x) \leq u(x) \quad \text{for} \quad x_0 \leq x \leq X. \tag{10.13}$$

The same conclusion is true if both D_+ are replaced by D^+ .

Proof. In order to be able to compare the derivatives $D_+ m$ and $D_+ u$ in (10.12), we consider points at which $m(x) = u(x)$. This is the main idea.

If (10.13) were not true, we could choose a point x_2 with $m(x_2) > u(x_2)$ and look for the first point x_1 to the left of x_2 with $m(x_1) = u(x_1)$. Then for small $h > 0$ we would have

$$\frac{m(x_1 + h) - m(x_1)}{h} > \frac{u(x_1 + h) - u(x_1)}{h}$$

and, by taking limits, $D_+ m(x_1) \geq D_+ u(x_1)$. This, however, contradicts (a) and (b), which give

$$D_+ m(x_1) \leq g(x_1, m(x_1)) = g(x_1, u(x_1)) < D_+ u(x_1). \quad \square$$

Many variant forms of this theorem are possible, for example by using left Dini derivatives (Walter 1970, Chap. II, §8, Theorem V).

Theorem 10.2 (The “fundamental lemma”). *Suppose that $y(x)$ is a solution of the system of differential equations $y' = f(x, y)$, $y(x_0) = y_0$, and that $v(x)$ is an approximate solution. If*

$$\begin{aligned} a) \quad & \|v(x_0) - y(x_0)\| \leq \varrho \\ b) \quad & \|v'(x + 0) - f(x, v(x))\| \leq \varepsilon \\ c) \quad & \|f(x, v) - f(x, y)\| \leq L\|v - y\|, \end{aligned}$$

then, for $x \geq x_0$, we have the error estimate

$$\|y(x) - v(x)\| \leq \varrho e^{L(x-x_0)} + \frac{\varepsilon}{L} (e^{L(x-x_0)} - 1). \tag{10.14}$$

Remark. The two terms in (10.14) express, respectively, the influence of the error ϱ in the initial values and the influence of the defect ε to the error of the approximate solution. It implies that the error depends continuously on both, and that for $\varrho = \varepsilon = 0$ we have $y(x) = v(x)$, i.e., uniqueness of the solution.

Proof. We put $m(x) = \|y(x) - v(x)\|$ and obtain, as in (10.7),

$$D_+ m(x) \leq L \cdot m(x) + \varepsilon, \quad m(x_0) \leq \varrho.$$

We shall try to compare this with the differential equation

$$u' = Lu + \varepsilon, \quad u(x_0) = \varrho. \quad (10.15)$$

Theorem 10.1 is not directly applicable. We therefore replace in (10.15) ε by $\varepsilon + \eta$, $\eta > 0$ and solve instead

$$u' = Lu + \varepsilon + \eta > Lu + \varepsilon, \quad u(x_0) = \varrho.$$

Now Theorem 10.1 gives the estimate (10.14) with ε replaced by $\varepsilon + \eta$. Since this estimate is true for *all* $\eta > 0$, it is also true for $\eta = 0$. \square

Variant form of Theorem 10.2. *The conditions*

- a) $\|v(x_0) - y(x_0)\| \leq \varrho$
- b) $\|v'(x+0) - f(x, v(x))\| \leq \delta(x)$
- c) $\|f(x, v) - f(x, y)\| \leq \ell(x)\|v - y\|$

imply for $x \geq x_0$

$$\|y(x) - v(x)\| \leq e^{L(x)} \left(\varrho + \int_{x_0}^x e^{-L(s)} \delta(s) ds \right), \quad L(x) = \int_{x_0}^x \ell(s) ds.$$

Proof. This is simply formula (3.3). \square

Theorem 10.3. *If the function $g(x, y)$ is continuous and satisfies a Lipschitz condition, then the implication (10.10) is true for continuous functions $m(x)$ and $u(x)$.*

Proof. Define functions $w_n(x)$, $v_n(x)$ by

$$\begin{aligned} w'_n(x) &= g(x, w_n(x)) + 1/n, & w_n(x_0) &= m(x_0), \\ v'_n(x) &= g(x, v_n(x)) - 1/n, & v_n(x_0) &= u(x_0), \end{aligned}$$

so that from Theorem 10.1

$$m(x) \leq w_n(x), \quad v_n(x) \leq u(x) \quad \text{for} \quad x_0 \leq x \leq X. \quad (10.16)$$

It follows from Theorem 10.2 that the functions $w_n(x)$ and $v_n(x)$ converge for $n \rightarrow \infty$ to the solutions of

$$\begin{aligned} w'(x) &= g(x, w(x)), & w(x_0) &= m(x_0), \\ v'(x) &= g(x, v(x)), & v(x_0) &= u(x_0), \end{aligned}$$

since the defect is $\pm 1/n$. Finally, because of $m(x_0) \leq u(x_0)$ and uniqueness we have $w(x) \leq v(x)$. Taking the limit $n \rightarrow \infty$ in (10.16) thus gives $m(x) \leq u(x)$. \square

A further generalization of Theorem 10.2 is possible if the Lipschitz condition (c) is replaced by something nonlinear such as

$$\|f(x, v) - f(x, y)\| \leq \omega(x, \|v - y\|).$$

Then the differential inequality for the error $m(x)$ is to be compared with the solution of

$$u' = \omega(x, u) + \delta(x) + \eta, \quad u(x_0) = \varrho, \quad \eta > 0.$$

See Walter (1970), Chap. II, §11 for more details.

Estimates Using One-Sided Lipschitz Conditions

As we already observed in Exercise 2 of I.9, and as has been known for a long time, much information about the errors can be lost by the use of positive Lipschitz constants L (e.g. (9.11), (9.11'), or (9.11'')) in the estimates (7.16), (7.17), or (7.18). The estimates all grow exponentially with x , even if the solutions and errors decay. Therefore many efforts have been made to obtain better error estimates, as for example the papers Eltermann (1955), Uhlmann (1957), Dahlquist (1959), and the references therein. We follow with great pleasure the particularly clear presentation of Dahlquist.

Let us estimate the derivative of $m(x) = \|v(x) - y(x)\|$ with more care than we did in (10.5): for $h > 0$ we have

$$\begin{aligned} m(x+h) &= \|v(x+h) - y(x+h)\| \\ &= \|v(x) - y(x) + h(v'(x) - y'(x))\| + \mathcal{O}(h^2) \\ &\leq \left\| v(x) - y(x) + h \left(f(x, v(x)) - f(x, y(x)) \right) \right\| + h\delta(x) + \mathcal{O}(h^2) \end{aligned} \quad (10.17)$$

by the use of (10.6) and (9.7). Here, we apply the mean value theorem to the function $y + hf(x, y)$ and obtain

$$m(x+h) \leq \left(\max_{\eta \in [y(x), v(x)]} \left\| I + h \frac{\partial f}{\partial y}(x, \eta) \right\| \right) \cdot m(x) + h\delta(x) + \mathcal{O}(h^2)$$

and finally for $h > 0$,

$$\frac{m(x+h) - m(x)}{h} \leq \max_{\eta \in [y(x), v(x)]} \frac{\left\| I + h \frac{\partial f}{\partial y}(x, \eta) \right\| - 1}{h} m(x) + \delta(x) + \mathcal{O}(h). \quad (10.18)$$

The expression on the right hand side of (10.18) leads us to the following definition:

Definition 10.4. Let Q be a square matrix, then we call

$$\mu(Q) = \lim_{h \rightarrow 0, h > 0} \frac{\|I + hQ\| - 1}{h} \quad (10.19)$$

the *logarithmic norm* of Q .

Here are formulas for its computation (Dahlquist (1959), p. 11, Eltermann (1955), p. 498, 499):

Theorem 10.5. *The logarithmic norm (10.19) is obtained by the following formulas: for the Euclidean norm (9.6),*

$$\mu(Q) = \lambda_{\max} = \text{largest eigenvalue of } \frac{1}{2}(Q^T + Q); \quad (10.20)$$

for the max-norm (9.6'),

$$\mu(Q) = \max_{k=1, \dots, n} \left(q_{kk} + \sum_{i \neq k} |q_{ki}| \right); \quad (10.20')$$

for the norm (9.6''),

$$\mu(Q) = \max_{i=1, \dots, n} \left(q_{ii} + \sum_{k \neq i} |q_{ki}| \right). \quad (10.20'')$$

Proofs. Formulas (10.20') and (10.20'') follow quite trivially from (9.11') and (9.11'') and the definition (10.19). The point is that the presence of I suppresses, for h sufficiently small, the absolute values for the diagonal elements. (10.20) is seen from the fact that the eigenvalues of

$$(I + hQ)^T (I + hQ) = I + h(Q^T + Q) + h^2 Q^T Q,$$

for $h \rightarrow 0$, converge to $1 + h\lambda_i$, where λ_i are the eigenvalues of $Q^T + Q$. \square

Remark. For complex-valued matrices the above formulas remain valid if one replaces Q by Q^* and q_{kk} , q_{ii} by $\operatorname{Re} q_{kk}$, $\operatorname{Re} q_{ii}$.

We now obtain from (10.18) the following improvement of Theorem 10.3.

Theorem 10.6. *Suppose that we have the estimates*

$$\mu\left(\frac{\partial f}{\partial y}(x, \eta)\right) \leq \ell(x) \quad \text{for} \quad \eta \in [y(x), v(x)] \quad \text{and} \quad (10.21)$$

$$\|v'(x+0) - f(x, v(x))\| \leq \delta(x), \quad \|v(x_0) - y(x_0)\| \leq \varrho.$$

Then for $x > x_0$ we have

$$\|y(x) - v(x)\| \leq e^{L(x)} \left(\varrho + \int_{x_0}^x e^{-L(s)} \delta(s) ds \right), \quad (10.22)$$

with $L(x) = \int_{x_0}^x \ell(s) ds$.

Proof. Since, for a fixed x , the segment $[v(x), y(x)]$ is compact,

$$K = \max_i \max_{[v(x), y(x)]} \left| \frac{\partial f_i}{\partial y_i} \right|$$

is finite. Then (see the proof of Theorem 10.5)

$$\frac{\|I + h \frac{\partial f}{\partial y}(x, \eta)\| - 1}{h} = \mu \left(\frac{\partial f}{\partial y}(x, \eta) \right) + \mathcal{O}(h)$$

where the $\mathcal{O}(h)$ -term is uniformly bounded in η . (For the norms (9.6') and (9.6'') this term is in fact zero for $h < 1/K$). Thus the condition (10.21) inserted into (10.18) gives

$$D_+ m(x) \leq \ell(x)m(x) + \delta(x).$$

Now the estimate (10.22) follows in the same way as that of Theorem 10.3. \square

Exercises

1. Apply Theorem 10.6 to the example of Exercise 2 of I.9. Observe the substantial improvement of the estimates.
2. Prove the following (a variant form of the famous "Gronwall lemma", Gronwall 1919): suppose that a positive function $m(x)$ satisfies

$$m(x) \leq \varrho + \varepsilon(x - x_0) + L \int_{x_0}^x m(s) ds =: w(x) \quad (10.23)$$

then

$$m(x) \leq \varrho e^{L(x-x_0)} + \frac{\varepsilon}{L} \left(e^{L(x-x_0)} - 1 \right); \quad (10.24)$$

- a) directly, by subtracting from (10.23)

$$u(x) = \varrho + \varepsilon(x - x_0) + L \int_{x_0}^x u(s) ds;$$

- b) by differentiating $w(x)$ in (10.23) and using Theorem 10.1.
 c) Prove Theorem 10.2 with the help of the above lemma of Gronwall. The same interrelations are, of course, also valid in more general situations.
3. Consider the problem $y' = \lambda y$, $y(0) = 1$ with $\lambda \geq 0$ and apply Euler's method with constant step size $h = 1/n$. Prove that

$$\frac{\lambda}{1 + \lambda/n} y_h(x) \leq D_+ y_h(x) \leq \lambda y_h(x)$$

and derive the estimate

$$\left(1 + \frac{\lambda}{n}\right)^n \leq e^\lambda \leq \left(1 + \frac{\lambda}{n}\right)^{n+\lambda} \quad \text{for } \lambda \geq 0.$$

4. Prove the following properties of the logarithmic norm:

- a) $\mu(\alpha Q) = \alpha \mu(Q)$ for $\alpha \geq 0$
- b) $-\|Q\| \leq \mu(Q) \leq \|Q\|$
- c) $\mu(Q + P) \leq \mu(Q) + \mu(P), \quad \mu\left(\int Q(t) dt\right) \leq \int \mu(Q(t)) dt$
- d) $|\mu(Q) - \mu(P)| \leq \|Q - P\|.$

5. For the Euclidean norm (10.20), $\mu(Q)$ is the smallest number satisfying

$$\langle v, Qv \rangle \leq \mu(Q) \|v\|^2.$$

This property is valid for all norms associated with a scalar product. Prove this.

6. Show that for the Euclidean norm the condition (10.21) is equivalent to

$$\langle y - z, f(x, y) - f(x, z) \rangle \leq \ell(x) \|y - z\|^2.$$

7. Observe, using an example of the form

$$y_1' = y_2, \quad y_2' = -y_1,$$

that a generalization of Theorem 10.1 to *systems* of first order differential equations, with inequalities interpreted component-wise, is not true in general (Müller 1926).

However, it is possible to prove such a generalization of Theorem 10.1 under the additional hypothesis that the functions $g_i(x, y_1, \dots, y_n)$ are *quasimonotone*, i.e., that

$$g_i(x, y_1, \dots, y_j, \dots, y_n) \leq g_i(x, y_1, \dots, z_j, \dots, y_n) \\ \text{if } y_j < z_j \quad \text{for all } j \neq i.$$

Try to prove this.

An important fact is that many systems from parabolic differential equations, such as equation (6.10), are quasimonotone. This allows many interesting applications of the ideas of this section (see Walter (1970), Chap. IV).

I.11 Systems of Linear Differential Equations

[Wronski] ... beschäftigte sich mit Mathematik, Mechanik und Physik, Himmelsmechanik und Astronomie, Statistik und politischer Ökonomie, mit Geschichte, Politik und Philosophie, ... er versuchte seine Kräfte in mehreren mechanischen und technischen Erfindungen. (S. Dickstein, III. Math. Kongr. 1904, p. 515)

With more knowledge about existence and uniqueness, and with more skill in linear algebra, we shall now, as did the mathematicians of the 19th century, better understand many points which had been left somewhat obscure in Sections I.4 and I.6 about linear differential equations of higher order.

Equation (4.9) divided by $a_n(x)$ (which is $\neq 0$ away from singular points) becomes

$$y^{(n)} + b_{n-1}(x)y^{(n-1)} + \dots + b_0(x)y = g(x), \quad b_i(x) = a_i(x)/a_n(x). \quad (11.1)$$

with $g(x) = f(x)/a_n(x)$. Introducing $y = y_1$, $y' = y_2, \dots, y^{(n-1)} = y_n$ we arrive at

$$\begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix} = \begin{pmatrix} 0 & 1 & & \\ 0 & 0 & \ddots & \\ \vdots & \vdots & \dots & 1 \\ -b_0(x) & -b_1(x) & \dots & -b_{n-1}(x) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ g(x) \end{pmatrix}. \quad (11.1')$$

We again denote by y the vector $(y_1, \dots, y_n)^T$ and by $f(x)$ the inhomogeneity, so that (11.1') becomes a special case of the following *system of linear differential equations*

$$y' = A(x)y + f(x), \quad (11.2)$$

$$A(x) = (a_{ij}(x)), \quad f(x) = (f_i(x)), \quad i, j = 1, \dots, n.$$

Here, the theorems of Section I.9 and I.10 apply without difficulty. Since the partial derivatives of the right hand side of (11.2) with respect to y_i are given by $a_{ki}(x)$, we have the Lipschitz estimate (see condition (c) of the variant form of Theorem 10.2), where $\ell(x) = \|A(x)\|$ in any subordinate matrix norm (9.11, 11', 11''). We apply Theorem 7.4, and the variant form of Theorem 10.2 with $v(x) = 0$ as "approximate solution". We may also take $\ell(x) = \mu(A(x))$ (see (10.20, 20', 20'')) and apply Theorem 10.6.

Theorem 11.1. *Suppose that $A(x)$ is continuous on an interval $[x_0, X]$. Then for any initial values $y_0 = (y_{10}, \dots, y_{n0})^T$ there exists for all $x_0 \leq x \leq X$ a unique*

solution of (11.2) satisfying

$$\|y(x)\| \leq e^{L(x)} \left(\|y_0\| + \int_{x_0}^x e^{-L(s)} \|f(s)\| ds \right) \quad (11.3)$$

$$L(x) = \int_{x_0}^x \ell(s) ds, \quad \ell(x) = \|A(x)\| \quad \text{or} \quad \ell(x) = \mu(A(x)).$$

For $f(x) \equiv 0$, $y(x)$ depends linearly on the initial values, i.e., there is a matrix $R(x, x_0)$ (the “resolvent”), such that

$$y(x) = R(x, x_0) y_0. \quad (11.4)$$

Proof. Since $\ell(x)$ is continuous and therefore bounded on any compact interval $[x_0, X]$, the estimate (11.3) shows that the solutions can be continued until the end. The linear dependence follows from the fact that, for $f \equiv 0$, linear combinations of solutions are again solutions, and from uniqueness. \square

Resolvent and Wronskian

From uniqueness we have that the solutions with initial values y_0 at x_0 and $y_1 = R(x_1, x_0) y_0$ at x_1 (see (11.4)) must be the same. Hence we have

$$R(x_2, x_0) = R(x_2, x_1) R(x_1, x_0) \quad (11.5)$$

for $x_0 \leq x_1 \leq x_2$. Finally by integrating backward from x_1, y_1 , i.e., by the co-ordinate transformation $x = x_1 - t$, $0 \leq t \leq x_1 - x_0$, we must arrive, again by uniqueness, at the starting values. Hence

$$R(x_0, x_1) = \left(R(x_1, x_0) \right)^{-1} \quad (11.6)$$

and (11.5) is true without any restriction on x_0, x_1, x_2 .

Let $y_i(x) = (y_{1i}(x), \dots, y_{ni}(x))^T$ (for $i = 1, \dots, n$) be a set of n solutions of the homogeneous differential equation

$$y' = A(x) y \quad (11.7)$$

which are *linearly independent* at $x = x_0$ (i.e., they form a *fundamental system*). We form the *Wronskian matrix* (Wronski 1810)

$$W(x) = \begin{pmatrix} y_{11}(x) & \dots & y_{1n}(x) \\ \vdots & & \vdots \\ y_{n1}(x) & \dots & y_{nn}(x) \end{pmatrix},$$

so that

$$W'(x) = A(x)W(x)$$

and all solutions can be written as

$$c_1 y_1(x) + \dots + c_n y_n(x) = W(x) c \quad \text{where} \quad c = (c_1, \dots, c_n)^T. \quad (11.8)$$

If this solution must satisfy the initial conditions $y(x_0) = y_0$, we obtain $c = W^{-1}(x_0)y_0$ and we have the formula

$$R(x, x_0) = W(x)W^{-1}(x_0). \quad (11.9)$$

Therefore all solutions are known if one has found n linearly independent solutions.

Inhomogeneous Linear Equations

Extending the idea of Joh. Bernoulli for (3.2) and Lagrange for (4.9), we now compute the solutions of the *inhomogeneous* equation (11.2) by letting c be “variable” in the “general solution” (11.8): $y(x) = W(x)c(x)$ (Liouville 1838). Exactly as in Section I.3 for (3.2) we obtain from (11.2) and (11.7) by differentiation

$$y' = W'c + Wc' = AWc + Wc' = AWc + f.$$

Hence $c' = W^{-1}f$. If we integrate this with integration constants c , we obtain

$$y(x) = W(x) \int_{x_0}^x W^{-1}(s)f(s) ds + W(x) c.$$

The initial conditions $y(x_0) = y_0$ imply $c = W^{-1}(x_0)y_0$ and we obtain:

Theorem 11.2 (“Variation of constants formula”). *Let $A(x)$ and $f(x)$ be continuous. Then the solution of the inhomogeneous equation $y' = A(x)y + f(x)$ satisfying the initial conditions $y(x_0) = y_0$ is given by*

$$\begin{aligned} y(x) &= W(x) \left(W^{-1}(x_0) y_0 + \int_{x_0}^x W^{-1}(s)f(s) ds \right) \\ &= R(x, x_0) y_0 + \int_{x_0}^x R(x, s)f(s) ds. \end{aligned} \quad (11.10)$$

The Abel-Liouville-Jacobi-Ostrogradskii Identity

We already know from (11.6) that $W(x)$ remains regular for all x . We now show that the *determinant* of $W(x)$ can be given explicitly as follows (Abel 1827, Liouville 1838, Jacobi 1845, §17):

$$\det(W(x)) = \det(W(x_0)) \cdot \exp\left(\int_{x_0}^x \operatorname{tr}(A(s)) ds\right), \quad (11.11)$$

$$\operatorname{tr}(A(x)) = a_{11}(x) + a_{22}(x) + \dots + a_{nn}(x)$$

which connects the determinant of $W(x)$ to the *trace* of $A(x)$.

For the *proof* of (11.11) (see also Exercise 2) we compute the derivative $\frac{d}{dx} \det(W(x))$. Since $\det(W(x))$ is multilinear, this derivative (by the Leibniz rule) is a sum of n terms, whose first is

$$T_1 = \det \begin{pmatrix} y'_{11} & y'_{12} & \cdots & y'_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nn} \end{pmatrix}.$$

We insert $y'_{1i} = a_{11}(x)y_{1i} + \dots + a_{1n}(x)y_{ni}$ from (11.7). All terms $a_{12}(x)y_{2i}, \dots, a_{1n}(x)y_{ni}$ disappear by subtracting multiples of lines 2 to n , so that $T_1 = a_{11}(x) \det(W(x))$. Summing all these terms we obtain finally

$$\frac{d}{dx} \det(W(x)) = (a_{11}(x) + \dots + a_{nn}(x)) \cdot \det(W(x)) \quad (11.12)$$

and (11.11) follows by integration. \square

Exercises

1. Compute the resolvent matrix $R(x, x_0)$ for the two systems

$$\begin{aligned} y'_1 &= y_1 & y'_1 &= y_2 \\ y'_2 &= 3y_2 & y'_2 &= -y_1 \end{aligned}$$

and check the validity of (11.5), (11.6) as well as (11.11).

2. Reconstruct Abel's original proof for (11.11), which was for the case

$$y''_1 + py'_1 + qy_1 = 0, \quad y''_2 + py'_2 + qy_2 = 0.$$

Multiply the equations by y_2 and y_1 respectively and subtract to eliminate q . Then integrate.

Use the result to obtain an identity for the two integrals

$$y_1(a) = \int_0^\infty e^{ax-x^2} x^{\alpha-1} dx, \quad y_2(a) = \int_0^\infty e^{-ax-x^2} x^{\alpha-1} dx,$$

which both satisfy

$$\frac{d^2 y_i}{da^2} - \frac{a}{2} \cdot \frac{dy_i}{da} - \frac{\alpha}{2} y_i = 0. \quad (11.13)$$

Hint. To verify (11.13), integrate from 0 to infinity the expression for $\frac{d}{dx}(\exp(ax - x^2)x^\alpha)$ (Abel 1827, case IV).

3. (Kummer 1839). Show that the general solution of the equation

$$y^{(n)}(x) = x^m y(x) \quad (11.14)$$

can be obtained by quadrature.

Hint. Differentiate (11.14) to obtain

$$y^{(n+1)} = x^m y' + m x^{m-1} y. \quad (11.15)$$

Suppose by recursion that the general solution of

$$\psi^{(n+1)} = x^{m-1} \psi, \quad \text{i.e.,} \quad \frac{d^{n+1}}{dx^{n+1}} \psi(xu) = x^{m-1} u^{m+n} \psi(xu) \quad (11.16)$$

is already known. Show that then

$$y(x) = \int_0^\infty u^{m-1} \exp\left(-\frac{u^{m+n}}{m+n}\right) \psi(xu) dx$$

is the general solution of (11.15), and, under some conditions on the parameters, also of (11.14). To simplify the computations, consider the function

$$g(u) = u^m \exp\left(-\frac{u^{m+n}}{m+n}\right) \psi(xu),$$

compute its derivative with respect to u , multiply by x^{m-1} , and integrate from 0 to infinity.

4. (Weak singularities for systems). Show that the linear system

$$y' = \frac{1}{x} (A_0 + A_1 x + A_2 x^2 + \dots) y \quad (11.17)$$

possesses solutions of the form

$$y(x) = x^q (v_0 + v_1 x + v_2 x^2 + \dots) \quad (11.18)$$

where v_0, v_1, \dots are vectors. Determine first q and v_0 , then recursively v_1, v_2 , etc. Observe that there exist n independent solutions of the form (11.18) if the eigenvalues of A_0 satisfy $\lambda_i \neq \lambda_j \pmod{\mathbb{Z}}$ (Fuchs 1866).

5. Find the general solution of the weakly singular systems

$$y' = \frac{1}{x} \begin{pmatrix} \frac{3}{4} & 1 \\ \frac{1}{4} & -\frac{1}{4} \end{pmatrix} y \quad \text{and} \quad y' = \frac{1}{x} \begin{pmatrix} \frac{3}{4} & 1 \\ -\frac{1}{4} & -\frac{1}{4} \end{pmatrix} y. \quad (11.19)$$

Hint. While the first is easy from Exercise 4, the second needs an additional idea (see formula (5.9)). A second possibility is to use the transformation $x = e^t$, $y(x) = z(t)$, and apply the methods of Section I.12.

I.12 Systems with Constant Coefficients

Die Technik der Integration der linearen Differentialgleichungen mit constanten Coeffizienten wird hier auf das Höchste entwickelt.
(F. Klein in Routh 1898)

Linearization

Systems of linear differential equations with constant coefficients form a class of equations for which the resolvent $R(x, x_0)$ can be computed explicitly. They generally occur by *linearization* of time-independent (i.e., *autonomous* or *permanent*) nonlinear differential equations

$$y'_i = f_i(y_1, \dots, y_n) \quad \text{or} \quad y''_i = f_i(y_1, \dots, y_n) \quad (12.1)$$

in the neighbourhood of a stationary point (Lagrange (1788), see also Routh (1860), Chap. IX, Thomson & Tait 1879). We choose the coordinates so that the stationary point under consideration is the origin, i.e., $f_i(0, \dots, 0) = 0$. We then expand f_i in its Taylor series and neglect all nonlinear terms:

$$y'_i = \sum_{k=1}^n \frac{\partial f_i}{\partial y_k}(0) y_k \quad \text{or} \quad y''_i = \sum_{k=1}^n \frac{\partial f_i}{\partial y_k}(0) y_k. \quad (12.1')$$

This is a system of equations with constant coefficients, as introduced in Section I.6 (see (6.4), (6.11)),

$$y' = Ay \quad \text{or} \quad y'' = Ay. \quad (12.1'')$$

Autonomous systems are invariant under a *shift* $x \rightarrow x + C$. We may therefore always assume that $x_0 = 0$. For arbitrary x_0 the resolvent is given by

$$R(x, x_0) = R(x - x_0, 0). \quad (12.2)$$

Diagonalization

We have seen in Section I.6 that the assumption $y(x) = v \cdot e^{\lambda x}$ leads to

$$Av = \lambda v \quad \text{or} \quad Av = \lambda^2 v, \quad (12.3)$$

hence $v \neq 0$ must be an *eigenvector* of A and λ the corresponding *eigenvalue* (in the first case; a square root of the eigenvalue in the second case, which we do not

consider any longer). From (12.3) we obtain by subtraction that there exists such a $v \neq 0$ if and only if the determinant

$$\chi_A(\lambda) := \det(\lambda I - A) = (\lambda - \lambda_1)(\lambda - \lambda_2) \dots (\lambda - \lambda_n) = 0. \quad (12.4)$$

This determinant is called the *characteristic polynomial of A*.

Suppose now that for the n eigenvalues λ_i the n eigenvectors v_i can be chosen linearly independent. We then have from (12.3)

$$A(v_1, v_2, \dots, v_n) = (v_1, v_2, \dots, v_n) \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_n),$$

or, if T is the matrix whose columns are the eigenvectors of A ,

$$T^{-1}AT = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_n). \quad (12.5)$$

On comparing (12.5) with (12.1''), we see that the differential equation simplifies considerably if we use the coordinate transformation

$$y(x) = Tz(x), \quad y'(x) = Tz'(x) \quad (12.6)$$

which leads to

$$z'(x) = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)z(x). \quad (12.7)$$

Thus the original system of differential equations decomposes into n single equations which are readily integrated to give

$$z(x) = \operatorname{diag}(\exp(\lambda_1 x), \exp(\lambda_2 x), \dots, \exp(\lambda_n x))z_0,$$

from which (12.6), yields

$$y(x) = T \operatorname{diag}(\exp(\lambda_1 x), \exp(\lambda_2 x), \dots, \exp(\lambda_n x))T^{-1}y_0. \quad (12.8)$$

The Schur Decomposition

Der Beweis ist leicht zu erbringen.

(Schur 1909)

The foregoing theory, beautiful as it may appear, has several drawbacks:

- a) Not all $n \times n$ matrices have a set of n linearly independent eigenvectors;
- b) Even if it is invertible, the matrix T can behave very badly (see Exercise 1).

However, for *symmetric* matrices a classical theory tells that A can always be diagonalized by orthogonal transformations. Let us therefore, with Schur (1909), extend this classical theory to non-symmetric matrices. A real matrix Q is called *orthogonal* if its column vectors are mutually orthogonal and of norm 1, i.e., if $Q^T Q = I$ or $Q^T = Q^{-1}$. A complex matrix Q is called *unitary* if $Q^* Q = I$ or $Q^* = Q^{-1}$, where Q^* is the *adjoint* matrix of Q , i.e., transposed and complex conjugate.

Theorem 12.1. a) (Schur 1909). *For each complex matrix A there exists a unitary matrix Q such that*

$$Q^*AQ = \begin{pmatrix} \lambda_1 & \times & \times & \dots & \times \\ & \lambda_2 & \times & \dots & \times \\ & & \ddots & & \vdots \\ & & & & \lambda_n \end{pmatrix}; \quad (12.9)$$

b) (Wintner & Murnaghan 1931). *For a real matrix A the matrix Q can be chosen real and orthogonal, if for each pair of conjugate eigenvalues $\lambda, \bar{\lambda} = \alpha \pm i\beta$ one allows the block*

$$\begin{pmatrix} \lambda & \times \\ & \bar{\lambda} \end{pmatrix} \quad \text{to be replaced by} \quad \begin{pmatrix} \times & \times \\ \times & \times \end{pmatrix}.$$

Proof. a) The matrix A has at least one eigenvector with eigenvalue λ_1 . We use this (normalized) vector as the first column of a matrix Q_1 . Its other columns are then chosen by arbitrarily completing the first one to an orthonormal basis. Then

$$AQ_1 = Q_1 \left(\begin{array}{c|ccc} \lambda_1 & \times & \dots & \times \\ 0 & & & A_2 \end{array} \right). \quad (12.10)$$

We then apply the same argument to the $(n-1)$ -dimensional matrix A_2 . This leads to

$$A_2\tilde{Q}_2 = \tilde{Q}_2 \left(\begin{array}{c|ccc} \lambda_2 & \times & \dots & \times \\ 0 & & & A_3 \end{array} \right).$$

With the unitary matrix

$$Q_2 = \left(\begin{array}{c|c} 1 & 0 \\ 0 & \tilde{Q}_2 \end{array} \right)$$

we obtain

$$Q_1^*AQ_1Q_2 = Q_2 \left(\begin{array}{cc|ccc} \lambda_1 & \times & & \times & \dots & \times \\ & \lambda_2 & & \times & \dots & \times \\ 0 & & & & & A_3 \end{array} \right).$$

A continuation of this process leads finally to a triangular matrix as in (12.9) with $Q = Q_1Q_2 \dots Q_{n-1}$.

b) Suppose A to be a real matrix. If λ_1 is real, Q_1 can be chosen real and orthogonal. Now let $\lambda_1 = \alpha + i\beta$ ($\beta \neq 0$) be a *non-real* eigenvalue with a corresponding eigenvector $u + iv$, i.e.,

$$A(u \pm iv) = (\alpha \pm i\beta)(u \pm iv) \quad (12.11)$$

or

$$Au = \alpha u - \beta v, \quad Av = \beta u + \alpha v. \quad (12.11')$$

Since $\beta \neq 0$, u and v are linearly independent. We choose an orthogonal basis \hat{u} , \hat{v} of the subspace spanned by u and v and take \hat{u} , \hat{v} as the first two columns of the orthogonal matrix Q_1 . We then have from (12.11')

$$AQ_1 = Q_1 \left(\begin{array}{cc|ccc} \times & \times & & \times & \dots & \times \\ \times & \times & & \times & \dots & \times \\ \hline 0 & & & & & A_3 \end{array} \right). \quad \square$$

Schur himself was not very proud of “his” decomposition, he just derived it as a tool for proving interesting properties of eigenvalues (see e.g., Exercise 2).

Clearly, if A is real and symmetric, $Q^T A Q$ will also be symmetric, and therefore diagonal (see also Exercise 3).

Numerical Computations

The above theoretical proof is still not of much practical use. It requires that one know the eigenvalues, but the computation of eigenvalues from the characteristic polynomial is one of the best-known stupidities of numerical analysis. Good numerical analysis turns it the other way round: the real matrix A is directly reduced, first to Hessenberg form, and then by a sequence of orthogonal transformations to the real Schur form of Wintner & Murnaghan (“QR-algorithm” of Francis, coded by Martin, Peters & Wilkinson, contribution II/14 in Wilkinson & Reinsch 1970). The eigenvalues then drop out. However, the produced code, called “HQR2”, does *not* give the Schur form of A , since it continues for the eigenvectors of A . Some manipulations must therefore be done to interrupt the code at the right moment (in the FORTRAN translation HQR2 of Eispack (1974), for example, the “340” of statement labelled “60” has to be replaced by “1001”). Happy “Matlab”-users just call “SCHUR”.

Whenever the Schur form has been obtained, the transformation $y(x) = Qz(x)$, $y'(x) = Qz'(x)$ (see (12.6)) leads to

$$\begin{pmatrix} z'_1 \\ \vdots \\ z'_{n-1} \\ z'_n \end{pmatrix} = \begin{pmatrix} \lambda_1 & b_{12} & \dots & b_{1,n-1} & b_{1n} \\ & \ddots & & \vdots & \vdots \\ & & & \lambda_{n-1} & b_{n-1,n} \\ & & & & \lambda_n \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ z_{n-1} \\ z_n \end{pmatrix}. \quad (12.12)$$

The last equation of this system is $z'_n = \lambda_n z_n$, and it can be integrated to give $z_n = \exp(\lambda_n x) z_{n0}$. Next, the equation for z_{n-1} is

$$z'_{n-1} = \lambda_{n-1} z_{n-1} + b_{n-1,n} z_n \quad (12.12')$$

with z_n known. This is a linear equation (inhomogeneous, if $b_{n-1,n} \neq 0$) which can be solved by Euler’s technique (Section I.4). Two different cases arise:

- a) If $\lambda_{n-1} \neq \lambda_n$ we put $z_{n-1} = E \exp(\lambda_{n-1}x) + F \exp(\lambda_n x)$, insert into (12.12') and compare coefficients. This gives $F = b_{n-1,n} z_{n0} / (\lambda_n - \lambda_{n-1})$ and $E = z_{n-1,0} - F$.
- b) If $\lambda_{n-1} = \lambda_n$ we set $z_{n-1} = (E + Fx) \exp(\lambda_n x)$ and obtain $F = b_{n-1,n} z_{n0}$ and $E = z_{n-1,0}$.

The next stage, following the same ideas, gives z_{n-2} , etc. Simple recursive formulas for the elements of the resolvent, which work in the case $\lambda_i \neq \lambda_j$, are obtained as follows (Parlett 1976): we assume

$$z_i(x) = \sum_{j=i}^n E_{ij} \exp(\lambda_j x) \quad (12.13)$$

and insert this into (12.12). After comparing coefficients, we obtain for $i = n, n-1, n-2$, etc.

$$E_{ik} = \frac{1}{\lambda_k - \lambda_i} \left(\sum_{j=i+1}^k b_{ij} E_{jk} \right), \quad k = i+1, i+2, \dots \quad (12.13')$$

$$E_{ii} = z_{i0} - \sum_{j=i+1}^n E_{ij}.$$

The Jordan Canonical Form

Simpler Than You Thought
(Amer. Math. Monthly 87 (1980) Nr. 9)

Whenever one is not afraid of badly conditioned matrices (see Exercise 1), and many mathematicians are not, the Schur form obtained above can be further transformed into the famous *Jordan canonical form*:

Theorem 12.2 (Jordan 1870, Livre deuxième, §5 and 6). *For every matrix A there exists a non-singular matrix T such that*

$$T^{-1}AT = \text{diag} \left\{ \begin{pmatrix} \lambda_1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & \\ & & & \lambda_1 \end{pmatrix}, \begin{pmatrix} \lambda_2 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & \\ & & & \lambda_2 \end{pmatrix}, \dots \right\}. \quad (12.14)$$

(The dimensions (≥ 1) of the blocks may vary and the λ_i are not necessarily distinct).

Proof. We may suppose that the matrix is already in the Schur form. This is of course possible in such a way that identical eigenvalues are grouped together on the principal diagonal.

The next step (see Fletcher & Sorensen 1983) is to remove all nonzero elements outside the upper-triangular blocks containing identical eigenvalues. We let

$$A = \begin{pmatrix} B & C \\ 0 & D \end{pmatrix}$$

where B and D are upper-triangular. The diagonal elements of B are all equal to λ_1 , whereas those of D are $\lambda_2, \lambda_3, \dots$ and all different from λ_1 . We search for a matrix S such that

$$\begin{pmatrix} B & C \\ 0 & D \end{pmatrix} \begin{pmatrix} I & S \\ 0 & I \end{pmatrix} = \begin{pmatrix} I & S \\ 0 & I \end{pmatrix} \begin{pmatrix} B & 0 \\ 0 & D \end{pmatrix}$$

or, equivalently,

$$BS + C = SD. \quad (12.15)$$

From this relation the matrix S can be computed column-wise as follows: the first column of (12.15) is $BS_1 + C_1 = \lambda_2 S_1$ (here S_j and C_j denote the j th column of S and C , respectively) which yields S_1 because λ_2 is not an eigenvalue of B . The second column of (12.15) yields $BS_2 + C_2 = \lambda_3 S_2 + d_{12} S_1$ and allows us to compute S_2 , etc.

In the following steps we treat each of the remaining blocks separately: we thus assume that all diagonal elements are equal to λ and transform the block recursively to the form stated in the theorem. Since $(A - \lambda I)^n = 0$ (n is the dimension of the matrix A) there exists an integer k ($1 \leq k \leq n$) such that

$$(A - \lambda I)^k = 0, \quad (A - \lambda I)^{k-1} \neq 0. \quad (12.16)$$

We fix a vector v such that $(A - \lambda I)^{k-1}v \neq 0$ and put

$$v_j = (A - \lambda I)^{k-j}v, \quad j = 1, \dots, k$$

so that

$$Av_1 = \lambda v_1, \quad Av_j = \lambda v_j + v_{j-1} \quad \text{for } j = 2, \dots, k.$$

The vectors v_1, \dots, v_k are linearly independent, because a multiplication of the expression $\sum_{j=1}^k c_j v_j = 0$ with $(A - \lambda I)^{k-1}$ yields $c_k = 0$, then a multiplication with $(A - \lambda I)^{k-2}$ yields $c_{k-1} = 0$, etc. As in the proof of the Schur decomposition (Theorem 12.1) we complete v_1, \dots, v_k to a basis of \mathbb{C}^n in such a way that (with $V = (v_1, \dots, v_n)$)

$$AV = V \begin{pmatrix} J & C \\ 0 & D \end{pmatrix}, \quad J = \left(\begin{array}{ccc} \lambda & 1 & \\ & \ddots & 1 \\ & & \lambda \end{array} \right) \Bigg\}^k \quad (12.17)$$

where D is upper-triangular with λ on its diagonal.

Our next aim is to eliminate the nonzero elements of C in (12.17). In analogy to (12.15) it is natural to search for a matrix S such that $JS + C = SD$. Unfortunately, such an S does not always exist because the eigenvalues of J and of D are

the same. However, it is possible to find S such that all elements of C are removed with the exception of its last line, i.e.,

$$\begin{pmatrix} J & C \\ 0 & D \end{pmatrix} \begin{pmatrix} I & S \\ 0 & I \end{pmatrix} = \begin{pmatrix} I & S \\ 0 & I \end{pmatrix} \begin{pmatrix} J & e_k c^T \\ 0 & D \end{pmatrix} \quad (12.18)$$

or equivalently

$$JS + C = e_k c^T + SD,$$

where $e_k = (0, \dots, 0, 1)^T$ and $c^T = (c_1, \dots, c_{n-k})$. This can be seen as follows: the first column of this relation becomes $(J - \lambda I)S_1 + C_1 = c_1 e_k$. Its last component yields c_1 and the other components determine the 2nd to k th elements of S_1 . The first element of S_1 can arbitrarily be put equal to zero. Then we compute S_2 from $(J - \lambda I)S_2 + C_2 = c_2 e_k + d_{12}S_1$, etc. We thus obtain a matrix S (with vanishing first line) such that (12.18) holds.

We finally show that the assumption $(A - \lambda I)^k = 0$ implies $c = 0$ in (12.18). Indeed, a simple calculation yields

$$\begin{pmatrix} J - \lambda I & e_k c^T \\ 0 & D - \lambda I \end{pmatrix}^k = \begin{pmatrix} 0 & \hat{C} \\ 0 & 0 \end{pmatrix}$$

where the first row of \hat{C} is equal to the row-vector c^T .

We have thus transformed A to block-diagonal form with blocks J of (12.17) and D . The procedure can now be repeated with the lower-dimensional matrix D . The product of all the occurring transformation matrices is then the matrix T in (12.14). \square

Corollary 12.3. *For every matrix A and for every number $\varepsilon \neq 0$ there exists a non-singular matrix T (depending on ε) such that*

$$T^{-1}AT = \text{diag} \left\{ \begin{pmatrix} \lambda_1 & \varepsilon & \\ & \ddots & \varepsilon \\ & & \lambda_1 \end{pmatrix}, \begin{pmatrix} \lambda_2 & \varepsilon & \\ & \ddots & \varepsilon \\ & & \lambda_2 \end{pmatrix}, \dots \right\}. \quad (12.14')$$

Proof. Multiply equation (12.14) from the right by $D = \text{diag}(1, \varepsilon, \varepsilon^2, \varepsilon^3, \dots)$ and from the left by D^{-1} . \square

Numerical difficulties in determining the Jordan canonical form are described in Golub & Wilkinson (1976). There exist also several computer programs, for example the one described in Kågstöm & Ruhe (1980).

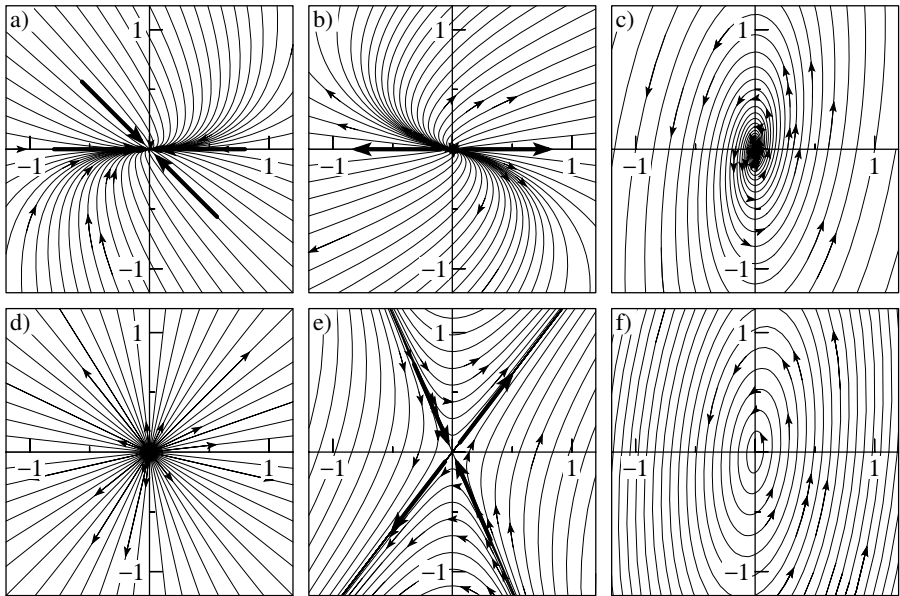
When the matrix A has been transformed to Jordan canonical form (12.14), the solutions of the differential equation $y' = Ay$ can be calculated by the method explained in (12.12'), case b):

$$y(x) = TDT^{-1}y_0 \quad (12.19)$$

where D is a block-diagonal matrix with blocks of the form

$$\begin{pmatrix} e^{\lambda x} & xe^{\lambda x} & \dots & \frac{x^k}{k!}e^{\lambda x} \\ & e^{\lambda x} & & \vdots \\ & & \ddots & xe^{\lambda x} \\ & & & e^{\lambda x} \end{pmatrix}$$

This is an extension of formula (12.8).



- | | | |
|---|---|---|
| a) $\begin{pmatrix} -1 & 1 \\ 0 & -2 \end{pmatrix}$ | b) $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ | c) $\begin{pmatrix} 1/3 & -1/3 \\ 2 & 0 \end{pmatrix}$ |
| d) $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ | e) $\begin{pmatrix} 1/3 & 1/3 \\ 1 & 0 \end{pmatrix}$ | f) $\begin{pmatrix} 1/6 & -1/3 \\ 2 & -1/6 \end{pmatrix}$ |

Fig. 12.1. Solutions of linear two dimensional systems

Geometric Representation

The geometric shapes of the solution curves of $y' = Ay$ are presented in Fig. 12.1 for dimension $n = 2$. They are plotted as paths in the phase-space (y_1, y_2) . The cases a), b), c) and e) are the linearized equations of (12.20) at the four critical points (see Fig. 12.2).

Much of this structure remains valid also for *nonlinear* systems (12.1) in the *neighbourhood of equilibrium points*. Exceptions may be “structurally unstable” cases such as complex eigenvalues with $\alpha = \text{Re}(\lambda) = 0$. This has been the subject of many papers discussing “critical points” or “singularities” (see e.g., the famous treatise of Poincaré (1881, 82, 85)).

In Fig. 12.2 we show solutions of the quadratic system

$$\begin{aligned} y_1' &= \frac{1}{3}(y_1 - y_2)(1 - y_1 - y_2) \\ y_2' &= y_1(2 - y_2) \end{aligned} \quad (12.20)$$

which possesses four critical points of all four possible structurally stable types (Exercise 4).

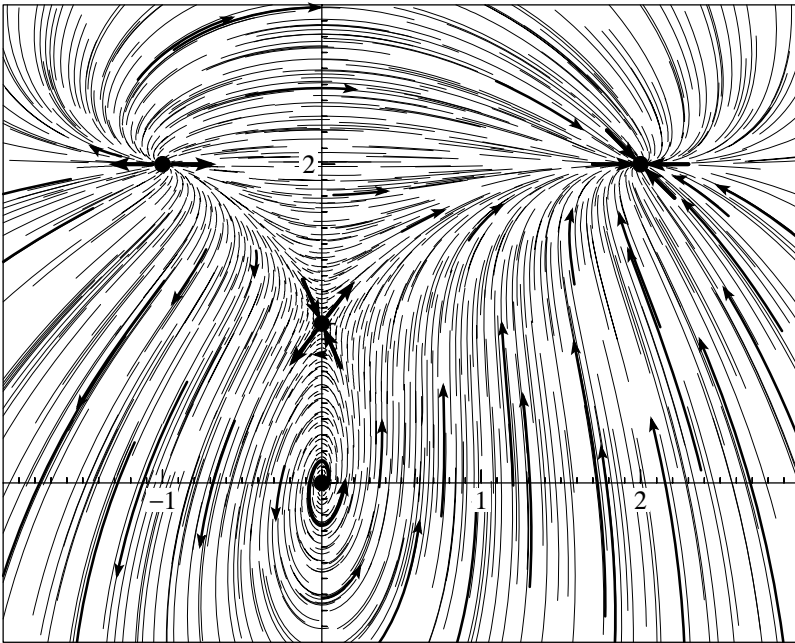


Fig. 12.2. Solution flow of System (12.20)

Exercises

1. a) Compute the eigenvectors of the matrix

$$A = \begin{pmatrix} -1 & 20 & & & & \\ & -2 & 20 & & & \\ & & -3 & 20 & & \\ & & & \ddots & \ddots & \\ & & & & -19 & 20 \\ & & & & & -20 \end{pmatrix} \quad (12.21)$$

by solving $(A - \lambda_i I)v_i = 0$.

Result. $v_1 = (1, 0, \dots)^T$, $v_2 = (1, -1/20, 0, \dots)^T$, $v_3 = (1, -2/20, 2/400, 0, \dots)^T$, $v_4 = (1, -3/20, 6/400, -6/8000, 0, \dots)^T$, etc.

- b) Compute numerically the inverse of $T = (v_1, v_2, \dots, v_n)$ and determine its largest element (answer: 4.5×10^{12}). The matrix T is thus very badly conditioned.
- c) Compute numerically or analytically from (12.13) the solutions of

$$y' = Ay, \quad y_i(0) = 1, \quad i = 1, \dots, 20. \quad (12.22)$$

Observe the “hump” (Moler & Van Loan 1978): although all eigenvalues of A are negative, the solutions first grow enormously before decaying to zero. This is typical of non-symmetric matrices and is connected with the bad condition of T (see Fig. 12.3).

Result.

$$y_1 = -\frac{20^{19}}{19!} e^{-20x} + \frac{(1+20)20^{18}}{18!} e^{-19x} - \frac{(1+20+20^2/2!)20^{17}}{17!} e^{-18x} \pm \dots$$

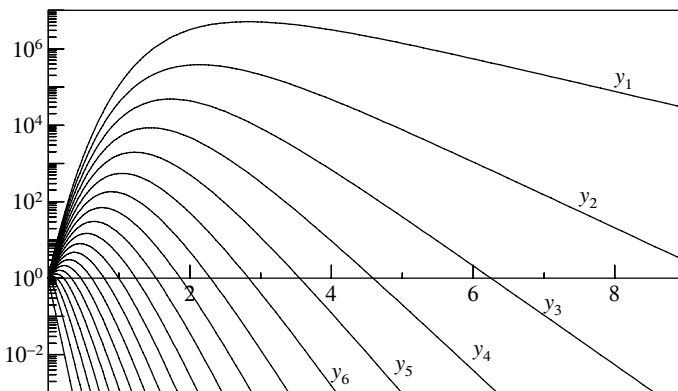


Fig. 12.3. Solutions of equation (12.22) with matrix (12.21)

2. (Schur). Prove that the eigenvalues of a matrix A satisfy the estimate

$$\sum_{i=1}^n |\lambda_i|^2 \leq \sum_{i,j=1}^n |a_{ij}|^2$$

and that equality holds iff A is orthogonally diagonalizable (see also Exercise 3).

Hint. $\sum_{i,j} |a_{ij}|^2$ is the trace of A^*A and thus invariant under unitary transformations Q^*AQ .

3. Show that the Schur decomposition $S = Q^*AQ$ is diagonal iff $A^*A = AA^*$. Such matrices are called *normal*. Examples are symmetric and skew-symmetric matrices.

Hint. The condition is equivalent to $S^*S = SS^*$.

4. Compute the four critical points of System (12.20), and for each of these points the eigenvalues and eigenvectors of the matrix $\partial f / \partial y$. Compare the results with Figs. 12.2 and 12.1.

5. Compute a Schur decomposition and the Jordan canonical form of the matrix

$$A = \frac{1}{9} \begin{pmatrix} 14 & 4 & 2 \\ -2 & 20 & 1 \\ -4 & 4 & 20 \end{pmatrix}.$$

Result. The Jordan canonical form is

$$\begin{pmatrix} 2 & 1 & \\ & 2 & \\ & & 2 \end{pmatrix}.$$

6. Reduce the matrices

$$A = \begin{pmatrix} \lambda & 1 & b & c \\ & \lambda & 1 & d \\ & & \lambda & 1 \\ & & & \lambda \end{pmatrix}, \quad A = \begin{pmatrix} \lambda & 1 & b & c \\ & \lambda & 0 & d \\ & & \lambda & 1 \\ & & & \lambda \end{pmatrix}$$

to Jordan canonical form. In the second case distinguish the possibilities $b + d = 0$ and $b + d \neq 0$.

I.13 Stability

The Examiners give notice that the following is the subject of the Prize to be adjudged in 1877: *The Criterion of Dynamical Stability*.
(S.G. Phear
(Vice-Chancellor), J. Challis, G.G. Stokes, J. Clerk Maxwell)

Introduction

“To illustrate the meaning of the question imagine a particle to slide down inside a smooth inclined cylinder along the lowest generating line, or to slide down outside along the highest generating line. In the former case a slight derangement of the motion would merely cause the particle to oscillate about the generating line, while in the latter case the particle would depart from the generating line altogether. The motion in the former case would be, in the sense of the question, stable, in the latter unstable . . . what is desired is, a corresponding condition enabling us to decide when a dynamically possible motion of a system is such, that *if slightly deranged the motion shall continue to be only slightly departed from*.” (“The Examiners” in Routh 1877).

Whenever no analytical solution of a problem is known, numerical solutions can only be obtained for specified initial values. But often one needs information about the stability behaviour of the solutions for all initial values in the neighbourhood of a certain equilibrium point. We again transfer the equilibrium point to the origin and define:

Definition 13.1. Let

$$y'_i = f_i(y_1, \dots, y_n), \quad i = 1, \dots, n \quad (13.1)$$

be a system with $f_i(0, \dots, 0) = 0$, $i = 1, \dots, n$. Then the origin is called *stable in the sense of Liapunov* if for any $\varepsilon > 0$ there is a $\delta > 0$ such that for the solutions, $\|y(x_0)\| < \delta$ implies $\|y(x)\| < \varepsilon$ for all $x > x_0$.

The first step, taken by Routh in his famous Adams Prize essay (Routh 1877), was to study the *linearized equation*

$$y'_i = \sum_{j=1}^n a_{ij} y_j, \quad a_{ij} = \frac{\partial f_i}{\partial y_j}(0). \quad (13.2)$$

(“The quantities x, y, z, \dots etc are said to be *small* when their squares can be neglected.”) From the general solution of (13.2) obtained in Section I.12, we immediately have

Theorem 13.1. *The linearized equation (13.2) is stable (in the sense of Liapunov) iff all roots of the characteristic equation*

$$\det(\lambda I - A) = a_0 \lambda^n + a_1 \lambda^{n-1} + \dots + a_{n-1} \lambda + a_n = 0 \quad (13.3)$$

satisfy $\operatorname{Re}(\lambda) \leq 0$, and the multiple roots, which give rise to Jordan chains, satisfy the strict inequality $\operatorname{Re}(\lambda) < 0$.

Proof. See (12.12) and (12.19). For Jordan chains the “secular” term (e.g., $E + Fx$ in the solution of (12.12), case (b)) which tends to infinity for increasing x , must be “killed” by an exponential with strictly negative exponent. \square

The Routh-Hurwitz Criterion

The next task, which leads to the famous Routh-Hurwitz criterion, was the verification of the conditions $\operatorname{Re}(\lambda) < 0$ directly from the coefficients of (13.3), without computing the roots. To solve this problem, Routh combined two known ideas: the first was Cauchy’s *argument principle*, saying that the number of roots of a polynomial $p(z) = u(z) + iv(z)$ inside a closed contour is equal to the number of (positive) rotations of the vector $(u(z), v(z))$, as z travels along the boundary in the positive sense (see e.g., Henrici (1974), p. 276). An example is presented in Fig. 13.1 for the polynomial

$$\begin{aligned} z^6 + 6z^5 + 16z^4 + 25z^3 + 24z^2 + 14z + 4 \\ = (z+1)(z+2)(z^2+z+1)(z^2+2z+2). \end{aligned} \quad (13.4)$$

On the half-circle $z = Re^{i\theta}$ ($\pi/2 \leq \theta \leq 3\pi/2$, R very large) the argument of $p(z)$, due to the dominant term z^n , makes $n/2$ positive rotations. In order to have all zeros of p in the negative half plane, we therefore need an additional $n/2$ positive rotations along the imaginary axis:

Lemma 13.2. *Let $p(z)$ be a polynomial of degree n and suppose that $p(iy) \neq 0$ for $y \in \mathbb{R}$. Then all roots of $p(z)$ are in the negative half-plane iff, along the imaginary axis, $\arg(p(iy))$ makes $n/2$ positive rotations for y from $-\infty$ to $+\infty$.* \square

The second idea was the use of Sturm’s theorem (Sturm 1829) which had its origin in Euclid’s algorithm for polynomials. Sturm made the discovery that in the division of the polynomial $p_{i-1}(y)$ by $p_i(y)$ it is better to take the remainder $p_{i+1}(y)$ with negative sign

$$p_{i-1}(y) = p_i(y)q_i(y) - p_{i+1}(y). \quad (13.5)$$

Then, due to the “Sturm sequence property”

$$\operatorname{sign}(p_{i+1}(y)) \neq \operatorname{sign}(p_{i-1}(y)) \quad \text{if} \quad p_i(y) = 0, \quad (13.6)$$

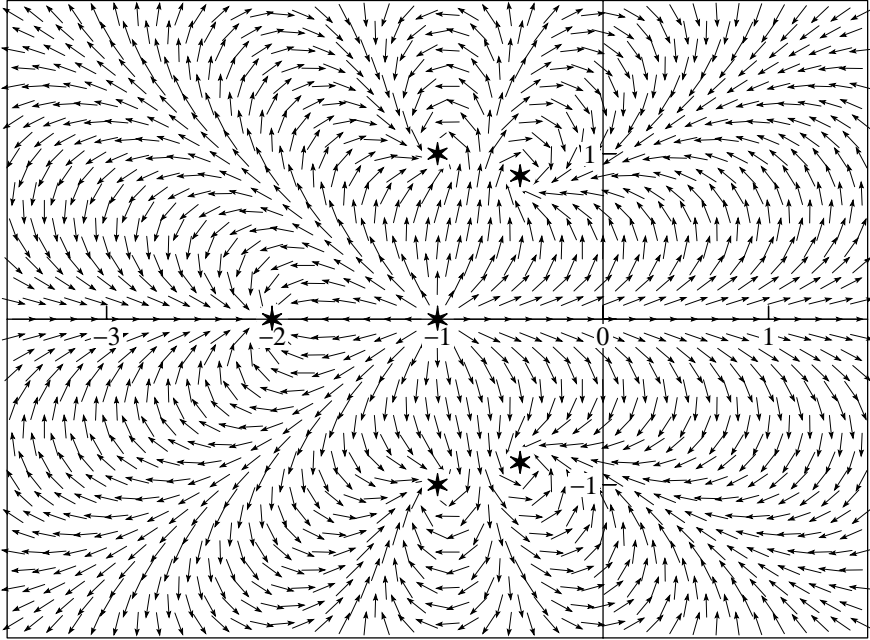


Fig. 13.1. Vector field of $\arg(p(z))$ for the polynomial $p(z)$ of (13.4)

the number of *sign changes*

$$w(y) = \text{No. of sign changes of } (p_0(y), p_1(y), \dots, p_m(y)) \quad (13.7)$$

does not vary at the zeros of $p_1(y), \dots, p_{m-1}(y)$. A consequence is the following

Lemma 13.3. *Suppose that a sequence $p_0(y), p_1(y), \dots, p_m(y)$ of real polynomials satisfies*

- i) $\deg(p_0) > \deg(p_1)$,
- ii) $p_0(y)$ and $p_1(y)$ not simultaneously zero,
- iii) $p_m(y) \neq 0$ for all $y \in \mathbb{R}$,
- iv) and the Sturm sequence property (13.6).

Then

$$\frac{w(\infty) - w(-\infty)}{2} \quad (13.8)$$

is equal to the number of rotations, measured in the positive direction, of the vector $(p_0(y), p_1(y))$ as y tends from $-\infty$ to $+\infty$.

Proof. Due to the Sturm sequence property, $w(y)$ does not change at zeros of $p_1(y), \dots, p_{m-1}(y)$. By assumption (iii) also $p_m(y)$ has no influence. Therefore $w(y)$ can change only at zeros of $p_0(y)$. If $w(y)$ increases by one at \hat{y} ,

either $p_0(y)$ changes from $+$ to $-$ and $p_1(\hat{y}) > 0$ or it changes from $-$ to $+$ and $p_1(\hat{y}) < 0$ ($p_1(\hat{y}) = 0$ is impossible by (ii)). In both situations the vector $(p_0(y), p_1(y))$ crosses the imaginary axis in the positive direction (see Fig. 13.2). If $w(y)$ decreases by one, $(p_0(y), p_1(y))$ crosses the imaginary axis in the negative direction. The result now follows from (i), since the vector $(p_0(y), p_1(y))$ is horizontal for $y \rightarrow -\infty$ and for $y \rightarrow +\infty$. \square

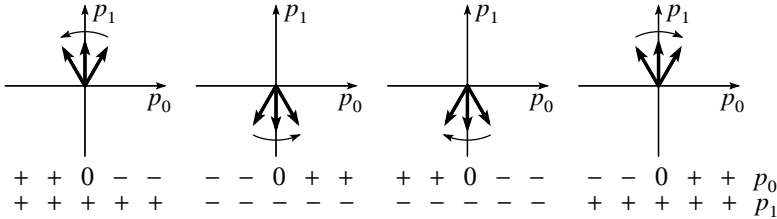


Fig. 13.2. Rotations of $(p_0(y), p_1(y))$ compared to $w(y)$

The two preceding lemmas together give us the desired criterion for stability: let the characteristic polynomial (13.3)

$$p(z) = a_0 z^n + a_1 z^{n-1} + \dots + a_n = 0, \quad a_0 > 0$$

be given. We divide $p(iy)$ by i^n and separate real and imaginary parts,

$$\begin{aligned} p_0(y) &= \operatorname{Re} \frac{p(iy)}{i^n} = a_0 y^n - a_2 y^{n-2} + a_4 y^{n-4} \pm \dots \\ p_1(y) &= -\operatorname{Im} \frac{p(iy)}{i^n} = a_1 y^{n-1} - a_3 y^{n-3} + a_5 y^{n-5} \pm \dots \end{aligned} \quad (13.9)$$

Due to the special structure of these polynomials, the Euclidean algorithm (13.5) is here particularly simple: we write

$$p_i(y) = c_{i0} y^{n-i} + c_{i1} y^{n-i-2} + c_{i2} y^{n-i-4} + \dots, \quad (13.10)$$

and have for the quotient in (13.5) $q_i(y) = (c_{i-1,0}/c_{i0})y$, provided that $c_{i0} \neq 0$. Now (13.10) inserted into (13.5) gives the following recursive formulas for the computation of the coefficients c_{ij} :

$$c_{i+1,j} = c_{i,j+1} \cdot \frac{c_{i-1,0}}{c_{i0}} - c_{i-1,j+1} = \frac{1}{c_{i0}} \det \begin{pmatrix} c_{i-1,0} & c_{i-1,j+1} \\ c_{i,0} & c_{i,j+1} \end{pmatrix}. \quad (13.11)$$

If $c_{i0} = 0$ for some i , the quotient $q_i(y)$ is a higher degree polynomial and the Euclidean algorithm stops at $p_m(y)$ with $m < n$.

The sequence $(p_i(y))$ obtained in this way obviously satisfies conditions (i) and (iv) of Lemma 13.3. Condition (ii) is equivalent to $p(iy) \neq 0$ for $y \in \mathbb{R}$, and (iii) is a consequence of (ii) since $p_m(y)$ is the *greatest common divisor* of $p_0(y)$ and $p_1(y)$.

Theorem 13.4 (Routh 1877). *All roots of the real polynomial (13.3) with $a_0 > 0$ lie in the negative half plane $\operatorname{Re} \lambda < 0$ if and only if*

$$c_{i0} > 0 \quad \text{for} \quad i = 0, 1, 2, \dots, n. \quad (13.12)$$

Remark. Due to the condition $c_{i0} > 0$, the division by c_{i0} in formula (13.11) can be omitted (common positive factor of $p_{i+1}(y)$), which leads to the same theorem (Routh (1877), p. 27: “. . . so that by remembering this simple cross-multiplication we may write down . . .”). This, however, is not advisable for n large because of possible overflow.

Proof. The coordinate systems (p_0, p_1) and $(\operatorname{Re}(p), \operatorname{Im}(p))$ are of *opposite* orientation. Therefore, $n/2$ positive rotations of $p(iy)$ correspond to $n/2$ negative rotations of $(p_0(y), p_1(y))$. If all roots of $p(\lambda)$ lie in the negative half plane $\operatorname{Re} \lambda < 0$, it follows from Lemmas 13.2 and 13.3 that $w(\infty) - w(-\infty) = -n$, which is only possible if $w(\infty) = 0$, $w(-\infty) = n$. This implies the positivity of all leading coefficients of $p_i(y)$.

On the other hand, if (13.12) is satisfied, we see that $p_n(y) \equiv c_{n0}$. Hence the polynomials $p_0(y)$ and $p_1(y)$ cannot have a common factor and $p(\lambda) \neq 0$ on the imaginary axis. We can now apply Lemmas 13.2 and 13.3 again to obtain the result. \square

Table 13.1.

Routh tableau for (13.4)

	$j = 0$	$j = 1$	$j = 2$	$j = 3$
$i = 0$	1	-16	24	-4
$i = 1$	6	-25	14	
$i = 2$	11.83	-21.67	4	
$i = 3$	14.01	-11.97		
$i = 4$	11.56	-4		
$i = 5$	7.12			
$i = 6$	4			

Table 13.2.

Routh tableau for (13.13)

	$j = 0$	$j = 1$	$j = 2$
$i = 0$	1	$-q$	s
$i = 1$	p	$-r$	
$i = 2$	$pq - r$	$-ps$	
$i = 3$	$(pq - r)r - p^2s$		
$i = 4$	$((pq - r)r - p^2s)ps$		

Example 1. The Routh tableau (13.11) for equation (13.4) is given in Table 13.1. It clearly satisfies the conditions for stability.

Example 2 (Routh 1877, p. 27). Express the stability conditions for the biquadratic $z^4 + pz^3 + qz^2 + rz + s = 0$.

$$(13.13)$$

The c_{ij} values (without division) are given in Table 13.2. We have stability iff

$$p > 0, \quad pq - r > 0, \quad (pq - r)r - p^2s > 0, \quad s > 0.$$

Computational Considerations

The actual computational use of Routh's criterion, in spite of its high historical importance and mathematical elegance, has two drawbacks for higher dimensions:

- 1) It is not easy to compute the characteristic polynomial for higher order matrices;
- 2) The use of the characteristic polynomial is very dangerous in the presence of rounding errors.

So, whenever one is not working with exact algebra or high precision, it is advisable to avoid the characteristic polynomial and use numerically stable algorithms for the eigenvalue problem (e.g., Eispack 1974).

Numerical experiments. 1. The $2n \times 2n$ dimensional matrix

$$A = \left(\begin{array}{ccc|ccc} -.05 & & & & -1 & \\ & \ddots & & & & \ddots \\ & & -.05 & & & -n \\ \hline 1 & & & -.05 & & \\ & \ddots & & & \ddots & \\ & & n & & & -.05 \end{array} \right)$$

has the characteristic polynomial

$$p(z) = \prod_{j=1}^n (z^2 + 0.1z + j^2 + 0.0025).$$

We computed the coefficients of p using double precision, and then applied the Routh algorithm in single precision (machine precision $= 6 \times 10^{-8}$). The results indicated stability for $n \leq 15$, but not for $n \geq 16$, although the matrix always has its eigenvalues $-0.05 \pm ki$ in the negative half plane. On the other hand, a direct computation of the eigenvalues of A with the use of Eispack subroutines gave no problem for any n .

2. We also tested the Routh algorithm at the (scaled) *numerators of the diagonal Padé approximations to $\exp(z)$*

$$1 + \frac{n}{2n}(nz) + \frac{n(n-1)}{(2n)(2n-1)} \frac{(nz)^2}{2!} + \frac{n(n-1)(n-2)}{(2n)(2n-1)(2n-2)} \frac{(nz)^3}{3!} + \dots, \quad (13.14)$$

which are also known to possess all zeros in \mathbb{C}^- . Here, the results were correct only for $n \leq 21$, and wrong for larger n due to rounding errors.

Liapunov Functions

We now consider the question whether the stability of the nonlinear system (13.1) “can really be determined by examination of the terms of the first order only” (Routh 1877, Chapt. VII). This theory, initiated by Routh and Poincaré, was brought to perfection in the famous work of Liapunov (1892). As a general reference to the enormous theory that has developed in the meantime we mention Rouche, Habets & Laloy (1977) and W. Hahn (1967).

Liapunov’s (and Routh’s) main tools are the so-called *Liapunov functions* $V(y_1, \dots, y_n)$, which should satisfy

$$\begin{aligned} V(y_1, \dots, y_n) &\geq 0, \\ V(y_1, \dots, y_n) &= 0 \quad \text{iff} \quad y_1 = \dots = y_n = 0 \end{aligned} \quad (13.15)$$

and along the solutions of (13.1)

$$\frac{d}{dx} V(y_1(x), \dots, y_n(x)) \leq 0. \quad (13.16)$$

Usually $V(y)$ behaves quadratically for small y and condition (13.15) means that

$$c\|y\|^2 \leq V(y) \leq C\|y\|^2, \quad C \geq c > 0. \quad (13.17)$$

The existence of such a Liapunov function is then a sufficient condition for stability of the origin.

We start with the *construction of a Liapunov function* for the linear case

$$y' = Ay. \quad (13.18)$$

This is best done in the basis which is naturally given by the *eigenvectors* (or Jordan chains) of A . We therefore introduce $y = Tz$, $z = T^{-1}y$, so that A is transformed to Jordan canonical form (12.14') $J = T^{-1}AT$ and (13.18) becomes

$$z' = Jz. \quad (13.19)$$

If we put

$$V_0(z) = \|z\|^2 \quad \text{and} \quad V(y) = V_0(T^{-1}y) = V_0(z), \quad (13.20)$$

the derivative of $V(y(x))$ becomes

$$\begin{aligned} \frac{d}{dx} V(y(x)) &= \frac{d}{dx} V_0(z(x)) = 2\operatorname{Re} \langle z(x), z'(x) \rangle \\ &= 2\operatorname{Re} \langle z(x), Jz(x) \rangle \leq 2\mu(J)V(y(x)). \end{aligned} \quad (13.21)$$

By (10.20) the logarithmic norm is given by

$$2\mu(J) = \text{largest eigenvalue of } (J + J^*).$$

The matrix $J + J^*$ is block-diagonal with tridiagonal blocks

$$\begin{pmatrix} 2 \operatorname{Re} \lambda_i & \varepsilon & & \\ \varepsilon & 2 \operatorname{Re} \lambda_i & \ddots & \\ & \ddots & \ddots & \varepsilon \\ & & \varepsilon & 2 \operatorname{Re} \lambda_i \end{pmatrix}. \quad (13.22)$$

Subtracting the diagonal and using formula (6.7a), we see that the eigenvalues of the m -dimensional matrix (13.22) are given by

$$2 \left(\operatorname{Re} \lambda_i + \varepsilon \cos \frac{\pi k}{m+1} \right), \quad k = 1, \dots, m. \quad (13.23)$$

As a consequence of this formula or by the use of Exercise 4 we have:

Lemma 13.5. *If all eigenvalues of A satisfy $\operatorname{Re} \lambda_i < -\varrho < 0$, then there exists a (quadratic) Liapunov function for equation (13.18) which satisfies*

$$\frac{d}{dx} V(y(x)) \leq -\varrho V(y(x)). \quad (13.24)$$

□

This last differential inequality implies that (Theorem 10.1)

$$V(y(x)) \leq V(y_0) \cdot \exp(-\varrho(x - x_0))$$

and ensures that $\lim_{x \rightarrow \infty} \|y(x)\| = 0$, i.e., *asymptotic stability*.

Stability of Nonlinear Systems

It is now easy to extend the same ideas to *nonlinear* equations. The following theorem is an example of such a result.

Theorem 13.6. *Let the nonlinear system*

$$y' = Ay + g(x, y) \quad (13.25)$$

be given with $\operatorname{Re} \lambda_i < -\varrho < 0$ for all eigenvalues of A . Further suppose that for each $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\|g(x, y)\| \leq \varepsilon \|y\| \quad \text{for} \quad \|y\| < \delta, \quad x \geq x_0. \quad (13.26)$$

Then the origin is (asymptotically) stable in the sense of Liapunov.

Proof. We use the Liapunov function $V(y)$ constructed for Lemma 13.5 and obtain from (13.25)

$$\frac{d}{dx} V(y(x)) \leq -\varrho V(y(x)) + 2 \operatorname{Re} \langle T^{-1} y(x), T^{-1} g(x, y(x)) \rangle. \quad (13.27)$$

Cauchy's inequality together with (13.26) yields

$$\frac{d}{dx}V(y(x)) \leq (-\varrho + \|T\| \cdot \|T^{-1}\|\varepsilon) \cdot V(y(x)). \quad (13.28)$$

For sufficiently small ε the right hand side is negative and we obtain asymptotic stability. \square

We see that, for nonlinear systems, stability *is only assured in a neighbourhood* of the origin. This can also be observed in Fig. 12.2. Another difference is that the *stability for eigenvalues on the imaginary axis can be destroyed*. An example for this (Routh 1877, pp. 95-96) is the system

$$y_1' = y_2, \quad y_2' = -y_1 + y_2^3. \quad (13.29)$$

Here, with the Liapunov function $V = (y_1^2 + y_2^2)/2$, we obtain $V' = y_2^4$ which is > 0 for $y_2 \neq 0$. Therefore all solutions with initial value $\neq 0$ increase. A survey of this question ("the center problem") together with its connection to limit cycles is given in Wanner (1983).

Stability of Non-Autonomous Systems

When the coefficients are not constant,

$$y' = A(x)y, \quad (13.30)$$

it is *not* a sufficient test of stability that the eigenvalues of A satisfy the conditions of stability for each instantaneous value of x .

Examples. 1. (Routh 1877, p. 96).

$$y_1' = y_2, \quad y_2' = -\frac{1}{4x^2}y_1 \quad (13.31)$$

which is satisfied by $y_1(x) = a\sqrt{x}$.

2. An example with eigenvalues strictly negative: we start with

$$B = \begin{pmatrix} -1 & 0 \\ 4 & -1 \end{pmatrix}, \quad y' = By.$$

An inspection of the derivative of $V = (y_1^2 + y_2^2)/2$ shows that V *increases* in the sector $2 - \sqrt{3} < y_2/y_1 < 2 + \sqrt{3}$. The idea is to take the initial value in this region and, for x increasing, to rotate the coordinate system with the same speed as the solution rotates:

$$y' = T(x)BT(-x)y = A(x)y, \quad T(x) = \begin{pmatrix} \cos ax & -\sin ax \\ \sin ax & \cos ax \end{pmatrix}. \quad (13.32)$$

For $y(0) = (1, 1)^T$, a good choice for a is $a = 2$ and (13.32) possesses the solution

$$y(x) = \left((\cos 2x - \sin 2x)e^x, (\cos 2x + \sin 2x)e^x \right)^T. \quad (13.33)$$

This solution is clearly unstable, while -1 remains for all x the double eigenvalue of $A(x)$. For more examples see Exercises 6 and 7 below.

We observe that stability theory for non-autonomous systems is more complicated. Among the cases in which stability can be shown are the following:

- 1) $a_{ii}(x) < 0$ and $A(x)$ is diagonally dominant; then $\mu(A(x)) \leq 0$ such that stability follows from Theorem 10.6.
- 2) $A(x) = B + C(x)$, with B constant and satisfying $\operatorname{Re} \lambda_i < -\rho < 0$ for its eigenvalues, and $\|C(x)\| < \varepsilon$ with ε so small that the proof of Theorem 13.6 can be applied.

Exercises

1. Express the stability conditions for the polynomials $z^2 + pz + q = 0$ and $z^3 + pz^2 + qz + r = 0$.

Result. a) $p > 0$ and $q > 0$; b) $p > 0$, $r > 0$ and $pq - r > 0$.

2. (Hurwitz 1895). Verify that condition (13.12) is equivalent to the positivity of the principal minors of the matrix

$$H = \begin{pmatrix} a_1 & a_3 & a_5 & \dots \\ a_0 & a_2 & a_4 & \dots \\ & a_1 & a_3 & \dots \\ & a_0 & a_2 & \dots \\ & & \dots & \dots \end{pmatrix} = \left(a_{2j-i} \right)_{i,j=1}^n$$

($a_k = 0$ for $k < 0$ and $k > n$). Understand that Routh's algorithm (13.11) is identical to a sort of Gaussian elimination transforming H to triangular form.

3. The polynomial

$$\frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6} \frac{z^5}{5!} + \frac{5 \cdot 4 \cdot 3 \cdot 2}{10 \cdot 9 \cdot 8 \cdot 7} \frac{z^4}{4!} + \frac{5 \cdot 4 \cdot 3}{10 \cdot 9 \cdot 8} \frac{z^3}{3!} + \frac{5 \cdot 4}{10 \cdot 9} \frac{z^2}{2!} + \frac{5}{10} z + 1$$

is the numerator of the $(5, 5)$ -Padé approximation to $\exp(z)$. Verify that all its roots satisfy $\operatorname{Re} z < 0$. Try to establish the result for general n (see e.g., Birkhoff & Varga (1965), Lemma 7).

4. (Gerschgorin). Prove that the eigenvalues of a matrix $A = (a_{ij})$ lie in the union of the discs

$$\left\{ z ; |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \right\}.$$

Hint. Write the formula $Ax = \lambda x$ in coordinates $\sum_j a_{ij}x_j = \lambda x_i$, put the diagonal elements on the right hand side and choose i such that $|x_i|$ is maximal.

5. Determine the stability of the origin for the system

$$\begin{aligned}y_1' &= -y_2 - y_1^2 - y_1y_2, \\y_2' &= y_1 + 2y_1y_2.\end{aligned}$$

Hint. Find a Liapunov function of degree 4 starting with $V = (y_1^2 + y_2^2)/2 + \dots$ such that $V' = K(y_1^2 + y_2^2)^2 + \dots$ and determine the sign of K .

6. (J. Lambert 1987). Consider the system

$$y' = A(x) \cdot y \quad \text{where} \quad A(x) = \begin{pmatrix} -1/4x & 1/x^2 \\ -1/4 & -1/4x \end{pmatrix}. \quad (13.34)$$

- a) Show that both eigenvalues of $A(x)$ satisfy $\operatorname{Re} \lambda < 0$ for all $x > 0$.
b) Compute $\mu(A)$ (from (10.20)) and show that

$$\mu(A) \leq 0 \quad \text{iff} \quad \sqrt{5} - 1 \leq x \leq \sqrt{5} + 1.$$

- c) Compute the general solution of (13.34).

Hint. Introduce the new functions $z_2(x) = y_2(x)$, $z_1(x) = xy_1(x)$ which leads to the second equation of (11.19) (Exercise 5 of Section I.11). The solution is

$$y_1(x) = x^{-3/4} \left(a + b \log x \right), \quad y_2(x) = x^{1/4} \left(-\frac{a}{2} + b \left(1 - \frac{1}{2} \log x \right) \right). \quad (13.35)$$

- d) Determine a and b such that $\|y(x)\|_2^2$ is *increasing* for $0 < x < \sqrt{5} - 1$.
e) Determine a and b such that $\|y(x)\|_2^2$ is *increasing* for $\sqrt{5} + 1 < x < \infty$.
Results. $b = 1.8116035 \cdot a$ for (d) and $b = 0.2462015 \cdot a$ for (e).

7. Find a counter-example for Fatou's conjecture

If $\ddot{y} + A(t)y = 0$ and $\forall t \quad 0 < C_1 \leq A(t) \leq C_2$ then $y(t)$ is stable (C.R. 189 (1929), p.967-969; for a solution see Perron (1930)).

8. Help James Watt (see original drawing from 1788 in Fig. 13.3) to solve the stability problem for his steam engine governor: if ω is the rotation speed of the engine, its acceleration is influenced by the steam supply and exterior work as follows:

$$\omega' = k \cos(\varphi + \alpha) - F, \quad k, F > 0.$$

Here α is a fixed angle and φ describes the motion of the governor. The acceleration of φ is determined by centrifugal force, weight, and friction as

$$\varphi'' = \omega^2 \sin \varphi \cos \varphi - g \sin \varphi - b\varphi', \quad g, b > 0.$$

Compute the equilibrium point $\varphi'' = \varphi' = \omega' = 0$ and determine under which conditions it is stable (the solution is easier for $\alpha = 0$).

Correct solutions should be sent to: James Watt, famous inventor of the steam engine, Westminster Abbey, 6HQ 1FX London.

Remark. Hurwitz' paper (1895) was motivated by a similar practical problem, namely "... die Regulierung von Turbinen des Badeortes Davos".

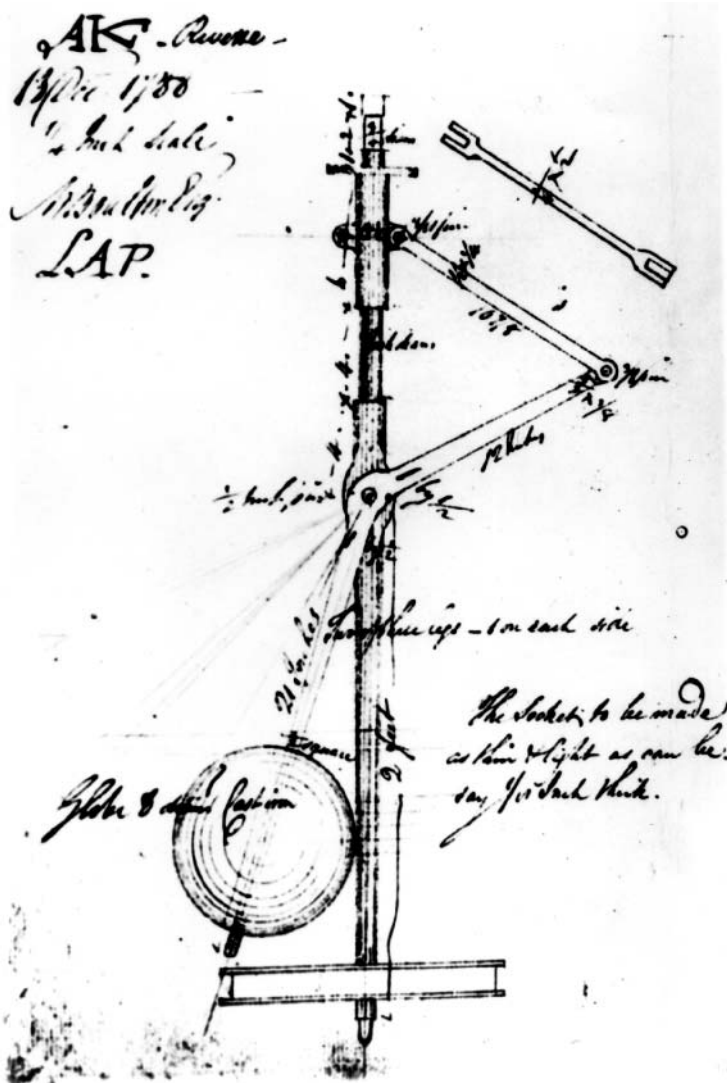


Fig. 13.3. James Watt's steam engine governor

I.14 Derivatives with Respect to Parameters and Initial Values

For a single equation, Dr. Ritt has solved the problem indicated in the title by a very simple and direct method . . . Dr. Ritt's proof cannot be extended immediately to a system of equations.

(T.H. Gronwall 1919)

In this section we consider the question whether the solutions of differential equations are differentiable

a) with respect to the initial values;

b) with respect to constant parameters in the equation;

and how these derivatives can be computed. Both questions are, of course, of extreme importance: once a solution has been computed (numerically) for given initial values, one often wants to know how small changes of these initial values affect the solutions. This question arises e.g. if some initial values are not known exactly and must be determined from other conditions, such as prescribed boundary values. Also, the initial values may contain errors, and the effect of these errors has to be studied. The same problems arise for unknown or wrong constant parameters in the defining equations.

Problems (a) and (b) are equivalent: let

$$y' = f(x, y, p), \quad y(x_0) = y_0 \quad (14.1)$$

be a system of differential equations containing a parameter p (or several parameters). We can add this parameter to the solutions

$$\begin{pmatrix} y' \\ p' \end{pmatrix} = \begin{pmatrix} f(x, y, p) \\ 0 \end{pmatrix}, \quad \begin{aligned} y(x_0) &= y_0 \\ p(x_0) &= p, \end{aligned} \quad (14.1')$$

so that the parameter becomes an initial value for $p' = 0$. Conversely, for a differential system

$$y' = f(x, y), \quad y(x_0) = y_0 \quad (14.2)$$

we can write $y(x) = z(x) + y_0$ and obtain

$$z' = f(x, z + y_0) = F(x, z, y_0), \quad z(x_0) = 0, \quad (14.2')$$

so that the initial value has become a parameter. Therefore, of the two problems (a) and (b), we start with (b) (as did Gronwall), because it seems simpler to us.

The Derivative with Respect to a Parameter

Usually, a given problem contains *several* parameters. But since we are interested in partial derivatives, we can treat one parameter after another while keeping the remaining ones fixed. It is therefore sufficient in the following theory to suppose that $f(x, y, p)$ depends only on *one* scalar parameter p .

When we replace the parameter p in (14.1) by q we obtain another solution, which we denote by $z(x)$:

$$z' = f(x, z, q), \quad z(x_0) = y_0. \quad (14.3)$$

It is then natural to subtract (14.1) from (14.3) and to linearize

$$\begin{aligned} z' - y' &= f(x, z, q) - f(x, y, p) \\ &= \frac{\partial f}{\partial y}(x, y, p)(z - y) + \frac{\partial f}{\partial p}(x, y, p)(q - p) + \varrho_1 \cdot (z - y) + \varrho_2 \cdot (q - p). \end{aligned} \quad (14.4)$$

If we put $(z(x) - y(x))/(q - p) = \psi(x)$ and drop the error terms, we obtain

$$\psi' = \frac{\partial f}{\partial y}(x, y(x), p)\psi + \frac{\partial f}{\partial p}(x, y(x), p), \quad \psi(x_0) = 0. \quad (14.5)$$

This equation is the key to the problem.

Theorem 14.1 (Gronwall 1919). *Suppose that for $x_0 \leq x \leq X$ the partial derivatives $\partial f/\partial y$ and $\partial f/\partial p$ exist and are continuous in the neighbourhood of the solution $y(x)$. Then the partial derivatives*

$$\frac{\partial y(x)}{\partial p} = \psi(x)$$

exist, are continuous, and satisfy the differential equation (14.5).

Proof. This theorem was the origin of the famous Gronwall lemma (see I.10, Exercise 2). We prove it here by the equivalent Theorem 10.2. Set

$$L = \max \left\| \frac{\partial f}{\partial y} \right\|, \quad A = \max \left\| \frac{\partial f}{\partial p} \right\| \quad (14.6)$$

where the \max is taken over the domain under consideration. When we consider $z(x)$ as an approximate solution for (14.1) we have for the defect

$$\|z'(x) - f(x, z(x), p)\| = \|f(x, z(x), q) - f(x, z(x), p)\| \leq A|q - p|,$$

therefore from Theorem 10.2

$$\|z(x) - y(x)\| \leq \frac{A}{L} |q - p| (e^{L(x-x_0)} - 1). \quad (14.7)$$

So for $|q - p|$ sufficiently small and $x_0 \leq x \leq X$, we can have $\|z(x) - y(x)\|$ arbitrarily small. By definition of differentiability and by (14.7), for each $\varepsilon > 0$

there is a δ such that the error terms in (14.4) satisfy

$$\|\varrho_1 \cdot (z - y) + \varrho_2 \cdot (q - p)\| \leq \varepsilon |q - p| \quad \text{if} \quad |q - p| \leq \delta. \quad (14.8)$$

(The situation is, in fact, a little more complicated: the δ for the bounds $\|\varrho_1\| < \varepsilon$ and $\|\varrho_2\| < \varepsilon$ may depend on x . But due to compactness and continuity, it can then be replaced by a uniform bound. Another possibility to overcome this little obstacle would be a bound on the second derivatives. But why should we worry about this detail? Gronwall himself did not mention it).

We now consider $(z(x) - y(x))/(q - p)$ as an approximate solution for (14.5) and apply Theorem 10.2 a second time. Its defect is by (14.8) and (14.4) bounded by ε and the linear differential equation (14.5) *also* has L as a Lipschitz constant (see (11.2)). Therefore from (10.14) we obtain

$$\left\| \frac{z(x) - y(x)}{q - p} - \psi(x) \right\| \leq \frac{\varepsilon}{L} (e^{L(x-x_0)} - 1)$$

which becomes arbitrarily small; this proves that $\psi(x)$ is the derivative of $y(x)$ with respect to p .

Continuity. The partial derivatives $\partial y / \partial p = \psi(x)$ are solutions of the differential equation (14.5), which we write as $\psi' = g(x, \psi, p)$, where by hypothesis g depends continuously on p . Therefore the continuous dependence of ψ on p follows again from Theorem 10.2. \square

Theorem 14.2. *Let $y(x)$ be the solution of equation (14.1) and consider the Jacobian*

$$A(x) = \frac{\partial f}{\partial y}(x, y(x), p). \quad (14.9)$$

Let $R(x, x_0)$ be the resolvent of the equation $y' = A(x)y$ (see (11.4)). Then the solution $z(x)$ of (14.3) with a slightly perturbed parameter q is given by

$$z(x) = y(x) + (q - p) \int_{x_0}^x R(x, s) \frac{\partial f}{\partial p}(s, y(s), p) ds + o(|q - p|) \quad (14.10)$$

Proof. This is the variation of constants formula (11.10) applied to (14.5). \square

It can be seen that the sensitivity of the solutions to changes of parameters is influenced firstly by the partial derivatives $\partial f / \partial p$ (which is natural), and secondly by the size of $R(x, s)$, i.e., by the stability of the differential equation with matrix (14.9).

Derivatives with Respect to Initial Values

Notation. We denote by $y(x, x_0, y_0)$ the solution $y(x)$ at the point x satisfying the initial values $y(x_0) = y_0$, and hope that no confusion arises from the use of the same letter y for two different functions.

The following identities are trivial by definition or follow from uniqueness arguments as for (11.6):

$$\frac{\partial y(x, x_0, y_0)}{\partial x} = f(x, y(x, x_0, y_0)) \quad (14.11)$$

$$y(x_0, x_0, y_0) = y_0 \quad (14.12)$$

$$y(x_2, x_1, y(x_1, x_0, y_0)) = y(x_2, x_0, y_0). \quad (14.13)$$

Theorem 14.3. *Suppose that the partial derivative of f with respect to y exists and is continuous. Then the solution $y(x, x_0, y_0)$ is differentiable with respect to y_0 and the derivative is given by the matrix*

$$\frac{\partial y(x, x_0, y_0)}{\partial y_0} = \Psi(x) \quad (14.14)$$

where $\Psi(x)$ is the resolvent of the so-called “variational equation”

$$\begin{aligned} \Psi'(x) &= \frac{\partial f}{\partial y}(x, y(x, x_0, y_0)) \cdot \Psi(x), \\ \Psi(x_0) &= I. \end{aligned} \quad (14.15)$$

Proof. We know from (14.2) and (14.2') that $\partial F/\partial z$ and $\partial F/\partial y_0$ are both equal to $\partial f/\partial y$, so the derivatives are known to exist by Theorem 14.1. In order to obtain formula (14.15), we just have to differentiate (14.11) and (14.12) with respect to y_0 . \square

We finally compute the derivative of $y(x, x_0, y_0)$ with respect to x_0 .

Theorem 14.4. *Under the same hypothesis as in Theorem 14.3, the solutions are also differentiable with respect to x_0 and the derivative is given by*

$$\frac{\partial y(x, x_0, y_0)}{\partial x_0} = -\frac{\partial y(x, x_0, y_0)}{\partial y_0} \cdot f(x_0, y_0). \quad (14.16)$$

Proof. Differentiate the identity

$$y(x_1, x_0, y(x_0, x_1, y_1)) = y_1,$$

which follows from (14.13), with respect to x_0 and apply (14.11) (see Exercise 1). \square

The Nonlinear Variation-of-Constants Formula

The following theorem is an extension of Theorem 11.2 to systems of non-linear differential equations.

Theorem 14.5 (Aleksseev 1961, Gröbner 1960). *Denote by y and z the solutions of*

$$y' = f(x, y), \quad y(x_0) = y_0, \quad (14.17a)$$

$$z' = f(x, z) + g(x, z), \quad z(x_0) = y_0, \quad (14.17b)$$

respectively and suppose that $\partial f / \partial y$ exists and is continuous. Then the solutions of (14.17a) and of the “perturbed” equation (14.17b) are connected by

$$z(x) = y(x) + \int_{x_0}^x \frac{\partial y}{\partial y_0}(x, s, z(s)) \cdot g(s, z(s)) ds. \quad (14.18)$$

Proof. We choose a subdivision $x_0 = s_0 < s_1 < s_2 < \dots < s_N = x$ (see Fig. 14.1). The descending curves represent the solutions of the unperturbed equation (14.17a) with initial values $s_i, z(s_i)$. The differences d_i are, due to the different slopes of $z(s)$ and $y(s)$ ((14.17b) minus (14.17a)), equal to $d_i = g(s_i, z(s_i)) \cdot \Delta s_i + o(\Delta s_i)$. This “error” at s_i is then “transported” to the final value x by the amount given in Theorem 14.3, to give

$$D_i = \frac{\partial y}{\partial y_0}(x, s_i, z(s_i)) \cdot g(s_i, z(s_i)) \cdot \Delta s_i + o(\Delta s_i). \quad (14.19)$$

Since $z(x) - y(x) = \sum_{i=1}^N D_i$, we obtain the integral in (14.18) after insertion of (14.19) and passing to the limit $\Delta s_i \rightarrow 0$. \square

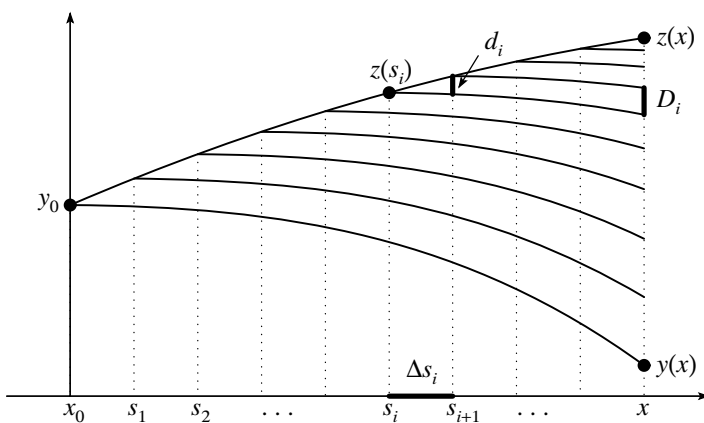


Fig. 14.1. Lady Windermere's fan, Act 2

If we also want to take into account a possible difference in the initial values, we may formulate:

Corollary 14.6. *Let $y(x)$ and $z(x)$ be the solutions of*

$$\begin{aligned} y' &= f(x, y), & y(x_0) &= y_0, \\ z' &= f(x, z) + g(x, z), & z(x_0) &= z_0, \end{aligned}$$

then

$$\begin{aligned} z(x) &= y(x) + \int_0^1 \frac{\partial y}{\partial y_0} \left(x, x_0, y_0 + s(z_0 - y_0) \right) \cdot (z_0 - y_0) ds \\ &\quad + \int_{x_0}^x \frac{\partial y}{\partial y_0} \left(x, s, z(s) \right) \cdot g(s, z(s)) ds. \end{aligned} \quad (14.20) \quad \square$$

These two theorems allow many estimates of the stability of general nonlinear systems. For *linear* systems, $\partial y / \partial y_0(x, s, z)$ is independent of z , and formulas (14.20) and (14.18) become the variation-of-constants formula (11.10). Also, by majorizing the integrals in (14.20) in a trivial way, one obtains the fundamental lemma (10.14) and also the variant form of Theorem 10.2.

Flows and Volume-Preserving Flows

Considérons des molécules fluides dont l'ensemble forme à l'origine des temps une certaine figure F_0 ; quand ces molécules se déplaceront, leur ensemble formera une nouvelle figure qui ira en se déformant d'une manière continue, et à l'instant t l'ensemble des molécules envisagées formera une nouvelle figure F .

(H. Poincaré, *Mécanique Céleste* 1899, Tome III, p.2)

We now turn our attention to a new interpretation of the Abel-Liouville-Jacobi-Ostrogradskii formula (11.11). Liouville and above all Jacobi (in his “Dynamik” 1843) used this formula extensively to obtain “first integrals”, i.e., relations between the solutions, so that the dimension of the system could be decreased and the analytic integration of the differential equations of mechanics becomes a little less hopeless. Poincaré then (see the quotation) introduced a much more geometric point of view: for an autonomous system of differential equations ¹

$$\frac{dy}{dt} = f(y) \quad (14.21)$$

we define the *flow* $\varphi_t : \mathbb{R}^n \rightarrow \mathbb{R}^n$ to be the function which associates, for a given t , to the initial value $y^0 \in \mathbb{R}^n$ the corresponding solution value at time t

$$\varphi_t(y^0) := y(t, 0, y^0). \quad (14.22)$$

¹ Due to the origin of these topics in Mechanics and Astronomy, we here use t for the independent variable.

For sets A of initial values we also study its behaviour under the action of the flow and write

$$\varphi_t(A) = \{y \mid y = y(t, 0, y^0), y^0 \in A\}. \quad (14.22')$$

We can imagine, with Poincaré, sets of “molecules” moving (and being deformed) with the flow.

Example 14.7. Fig. 14.2 shows, for the two-dimensional system (12.20) (see Fig. 12.2), the transformations which three sets A, B, C ² undergo when t passes from 0 to 0.2, 0.4 and (for C) 0.6. It can be observed that these sets quickly lose very much of their beauty.

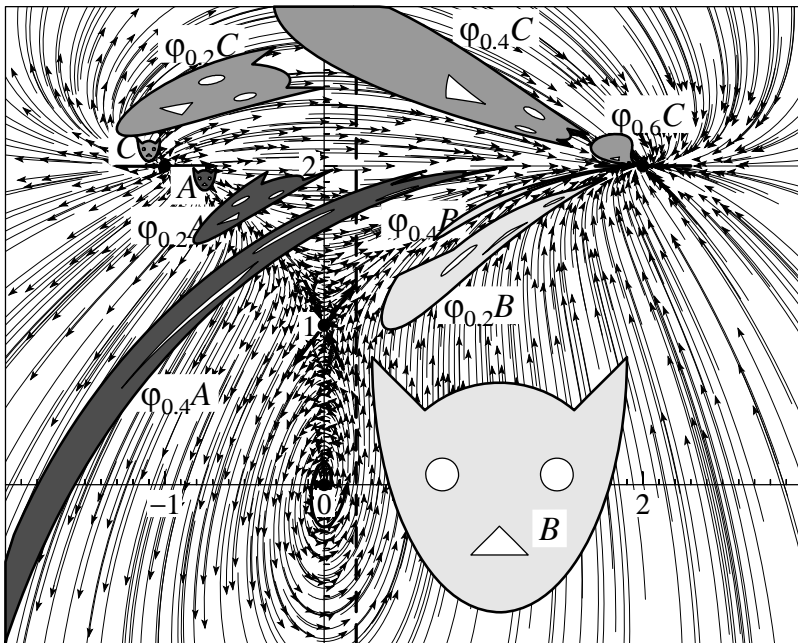


Fig. 14.2. Transformation of three sets under a flow

Now divide A into “infinitely small” cubes I of sides dy_1^0, \dots, dy_n^0 . The image $\varphi_t(I)$ of such a cube is an infinitely small parallelepiped. It is created by the columns of $\partial y / \partial y^0(t, 0, y^0)$ scaled by dy_i^0 , and its volume is $\det(\partial y / \partial y^0(t, 0, y^0)) \cdot dy_1^0 \dots dy_n^0$. Adding up all these volumes (over A) or, more precisely, using the transformation formula for multiple integrals

² The resemblance of these sets with a certain feline animal is not entirely accidental; we chose it in honour of V.I. Arnol'd.

(Euler 1769b, Jacobi 1841), we obtain

$$\text{Vol}(\varphi_t(A)) = \int_{\varphi_t(A)} dy = \int_A \left| \det \left(\frac{\partial y}{\partial y^0}(t, 0, y^0) \right) \right| dy^0.$$

Next we use formula (11.11) together with (14.15)

$$\det \left(\frac{\partial y}{\partial y^0}(t, 0, y^0) \right) = \exp \left(\int_0^t \text{tr} (f'(y(s, 0, y^0))) ds \right) \quad (14.23)$$

and we obtain

Theorem 14.8. *Consider the system (14.21) with continuously differentiable function $f(y)$.*

a) *For a set $A \subset \mathbb{R}^n$ the total volume of $\varphi_t(A)$ satisfies*

$$\text{Vol}(\varphi_t(A)) = \int_A \exp \left(\int_0^t \text{tr} (f'(y(s, 0, y^0))) ds \right) dy^0. \quad (14.24)$$

b) *If $\text{tr} (f'(y)) = 0$ along the solution, the flow is volume-preserving, i.e., $\text{Vol}(\varphi_t(A)) = \text{Vol}(A)$.* □

Example 14.9. For the system (12.20) we have

$$f'(y) = \begin{pmatrix} (1 - 2y_1)/3 & (2y_2 - 1)/3 \\ 2 - y_2 & -y_1 \end{pmatrix} \quad \text{and} \quad \text{tr} (f'(y)) = (1 - 5y_1)/3.$$

The trace of $f'(y)$ changes sign at the line $y_1 = 1/5$. To its left the volume increases, to the right we have decreasing volumes. This can clearly be seen in Fig. 14.2.

Example 14.10. For the mathematical pendulum (with y_1 the angle of deviation from the vertical)

$$\begin{aligned} \dot{y}_1 &= y_2 \\ \dot{y}_2 &= -\sin y_1 \end{aligned} \quad f'(y) = \begin{pmatrix} 0 & 1 \\ -\cos y_1 & 0 \end{pmatrix} \quad (14.25)$$

we have $\text{tr} (f'(y)) = 0$. Therefore the flow, although treating the cats quite badly, at least preserves their areas (Fig. 14.3).

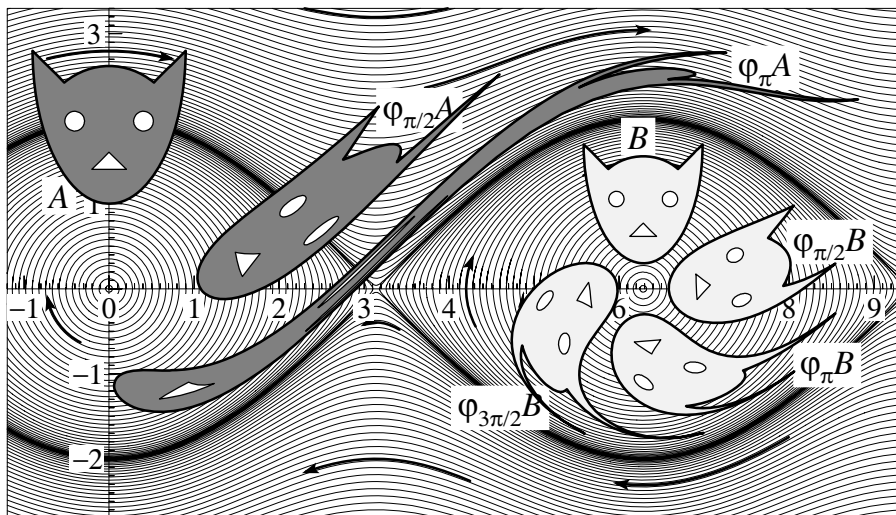


Fig. 14.3. Cats, beware of pendulums!

Canonical Equations and Symplectic Mappings

Let $H(p_1, \dots, p_n, q_1, \dots, q_n)$ be a twice continuously differentiable function of $2n$ variables and (see (6.26))

$$\dot{p}_i = -\frac{\partial H}{\partial q_i}(p, q), \quad \dot{q}_i = \frac{\partial H}{\partial p_i}(p, q) \quad (14.26)$$

the corresponding canonical system of differential equations. Small variations of the initial values lead to variations $\delta p_i(t), \delta q_i(t)$ of the solution of (14.26). By Theorem 14.3 (variational equation) these satisfy

$$\begin{aligned} \dot{\delta p}_i &= -\sum_{j=1}^n \frac{\partial^2 H}{\partial p_j \partial q_i}(p, q) \cdot \delta p_j - \sum_{j=1}^n \frac{\partial^2 H}{\partial q_j \partial q_i}(p, q) \cdot \delta q_j \\ \dot{\delta q}_i &= \sum_{j=1}^n \frac{\partial^2 H}{\partial p_j \partial p_i}(p, q) \cdot \delta p_j + \sum_{j=1}^n \frac{\partial^2 H}{\partial q_j \partial p_i}(p, q) \cdot \delta q_j. \end{aligned} \quad (14.27)$$

The upper left block of the Jacobian matrix is the negative transposed of the lower right block. As a consequence, the trace of the Jacobian of (14.27) is identically zero and the corresponding flow is volume-preserving ("Theorem of Liouville").

But there is much more than that (Poincaré 1899, vol. III, p. 43): consider a two-dimensional manifold A in the $2n$ -dimensional flow. We represent it as a (differentiable) map of a compact set $K \subset \mathbb{R}^2$ into \mathbb{R}^{2n} (Fig. 14.4)

$$\begin{aligned} \Phi : \quad K &\longrightarrow A \subset \mathbb{R}^{2n} \\ (u, v) &\longmapsto (p^0(u, v), q^0(u, v)) \end{aligned} \quad (14.28)$$

We let $\pi_i(A)$ be the projection of A onto the (p_i, q_i) -coordinate plane and consider the *sum of the oriented areas of $\pi_i(A)$* . We shall see that this is also an invariant.

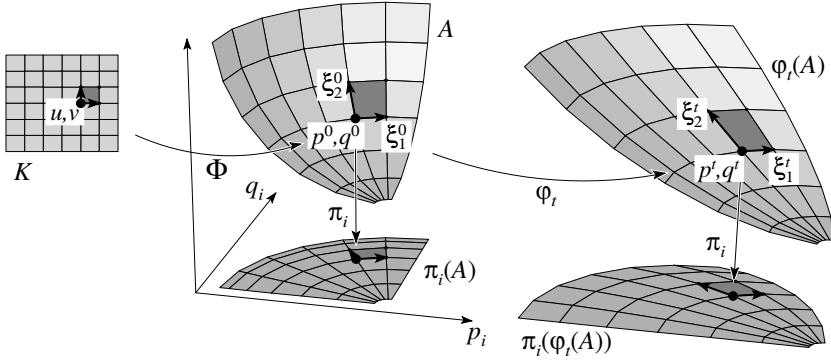


Fig. 14.4. Two-dimensional manifold in the flow

The oriented area of $\pi_i(A)$ is a surface integral over A which is defined, with the transformation formula in mind, as

$$\text{or.area}(\pi_i(A)) = \iint_K \det \begin{pmatrix} \frac{\partial p_i^0}{\partial u} & \frac{\partial p_i^0}{\partial v} \\ \frac{\partial q_i^0}{\partial u} & \frac{\partial q_i^0}{\partial v} \end{pmatrix} du dv. \quad (14.29)$$

For the computation of the area of $\pi_i(\varphi_t(A))$, after the action of the flow, we use the composition $\varphi_t \circ \Phi$ as coordinate map (Fig. 14.4). This produces, with p_i^t, q_i^t being the i th respectively $(n+i)$ th component of this map,

$$\text{or.area}(\pi_i(\varphi_t(A))) = \iint_K \det \begin{pmatrix} \frac{\partial p_i^t}{\partial u} & \frac{\partial p_i^t}{\partial v} \\ \frac{\partial q_i^t}{\partial u} & \frac{\partial q_i^t}{\partial v} \end{pmatrix} du dv. \quad (14.30)$$

There is no theoretical difficulty in differentiating this expression with respect to t and summing for $i = 1, \dots, n$. This will give zero and the invariance is established.

The proof, however, becomes more elegant if we introduce *exterior differential forms* (E. Cartan 1899). These, originally “expressions purement symboliques”, are today understood as *multilinear maps on the tangent space* (for more details see “Chapter 7” of Arnol’d 1974). In our case the one-forms dp_i , respectively dq_i , map a tangent vector ξ to its i th, respectively $(n+i)$ th, component. The *exterior product* $dp_i \wedge dq_i$ is a bilinear map acting on a pair of vectors

$$\begin{aligned} (dp_i \wedge dq_i)(\xi_1, \xi_2) &= \det \begin{pmatrix} dp_i(\xi_1) & dp_i(\xi_2) \\ dq_i(\xi_1) & dq_i(\xi_2) \end{pmatrix} \\ &= dp_i(\xi_1)dq_i(\xi_2) - dp_i(\xi_2)dq_i(\xi_1) \end{aligned} \quad (14.31)$$

and satisfies Grassmann's rules for exterior multiplication

$$dp_i \wedge dp_j = -dp_j \wedge dp_i, \quad dp_i \wedge dp_i = 0. \quad (14.32)$$

For the two tangent vectors (see Fig. 14.4)

$$\begin{aligned} \xi_1^0 &= \left(\frac{\partial p_1^0}{\partial u}(u, v), \dots, \frac{\partial p_n^0}{\partial u}(u, v), \frac{\partial q_1^0}{\partial u}(u, v), \dots, \frac{\partial q_n^0}{\partial u}(u, v) \right)^T \\ \xi_2^0 &= \left(\frac{\partial p_1^0}{\partial v}(u, v), \dots, \frac{\partial p_n^0}{\partial v}(u, v), \frac{\partial q_1^0}{\partial v}(u, v), \dots, \frac{\partial q_n^0}{\partial v}(u, v) \right)^T \end{aligned} \quad (14.33)$$

the expression (14.31) is precisely the integrand of (14.29). If we introduce the differential 2-form

$$\omega^2 = \sum_{i=1}^n dp_i \wedge dq_i \quad (14.34)$$

then our candidate for invariance becomes

$$\sum_{i=1}^n \text{or.area}(\pi_i(A)) = \iint_K \omega^2(\xi_1^0, \xi_2^0) du dv.$$

After the action of the flow we have the tangent vectors

$$\xi_1^t = \varphi'_t(p^0, q^0) \cdot \xi_1^0, \quad \xi_2^t = \varphi'_t(p^0, q^0) \cdot \xi_2^0$$

and

$$\sum_{i=1}^n \text{or.area}(\pi_i(\varphi_t(A))) = \iint_K \omega^2(\xi_1^t, \xi_2^t) du dv$$

(see (14.30)). We shall see that $\omega^2(\xi_1^t, \xi_2^t) = \omega^2(\xi_1^0, \xi_2^0)$.

Definition 14.11. For a differentiable function $g : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ we define the differential form $g^*\omega^2$ by

$$(g^*\omega^2)(\xi_1, \xi_2) := \omega^2(g'(p, q)\xi_1, g'(p, q)\xi_2). \quad (14.35)$$

Such a function g is called *symplectic* (a name suggested by H. Weyl 1939, p. 165) if

$$g^*\omega^2 = \omega^2, \quad (14.36)$$

i.e., if the 2-form ω^2 is invariant under g .

Theorem 14.12. *The flow of a canonical system (14.26) is symplectic, i.e.,*

$$(\varphi_t)^*\omega^2 = \omega^2 \quad \text{for all } t. \quad (14.37)$$

Proof. We compute the derivative of $\omega^2(\xi_1^t, \xi_2^t)$ (see (14.35)) with respect to t by

the Leibniz rule. This gives

$$\frac{d}{dt} \left(\sum_{i=1}^n (dp_i \wedge dq_i)(\xi_1^t, \xi_2^t) \right) = \sum_{i=1}^n (dp_i \wedge dq_i)(\dot{\xi}_1^t, \xi_2^t) + \sum_{i=1}^n (dp_i \wedge dq_i)(\xi_1^t, \dot{\xi}_2^t). \quad (14.38)$$

Since the vectors ξ_1^t and ξ_2^t satisfy the variational equation (14.27), we have

$$\begin{aligned} \frac{d}{dt} \omega^2(\xi_1^t, \xi_2^t) &= \sum_{i,j=1}^n \left(-\frac{\partial^2 H}{\partial p_j \partial q_i} dp_j \wedge dq_i - \frac{\partial^2 H}{\partial q_j \partial q_i} dq_j \wedge dq_i \right. \\ &\quad \left. + \frac{\partial^2 H}{\partial p_j \partial p_i} dp_i \wedge dp_j + \frac{\partial^2 H}{\partial q_j \partial p_i} dp_i \wedge dq_j \right) (\xi_1^t, \xi_2^t). \end{aligned} \quad (14.39)$$

The first and last terms in this formula cancel by symmetry of the partial derivatives. Further, the properties (14.32) imply that

$$\sum_{i,j=1}^n \frac{\partial^2 H}{\partial p_i \partial p_j}(p, q) dp_i \wedge dp_j = \sum_{i < j} \left(\frac{\partial^2 H}{\partial p_i \partial p_j}(p, q) - \frac{\partial^2 H}{\partial p_j \partial p_i}(p, q) \right) dp_i \wedge dp_j$$

vanishes. Since the last remaining term cancels in the same way, the derivative (14.38) vanishes identically. \square

Example 14.13. We use the spherical pendulum in canonical form (6.28)

$$\begin{aligned} \dot{p}_1 &= p_2^2 \frac{\cos q_1}{\sin^3 q_1} - \sin q_1 & \dot{p}_2 &= 0 \\ \dot{q}_1 &= p_1 & \dot{q}_2 &= \frac{p_2}{\sin^2 q_1} \end{aligned} \quad (14.40)$$

and for A the familiar two-dimensional cat placed in \mathbb{R}^4 such that its projection to (p_1, q_1) is a line; i.e., with zero area. It can be seen that with increasing t the area in (p_1, q_1) increases and the area in (p_2, q_2) decreases. Their sum remains constant. Observe that for larger t the left ear in (p_1, q_1) is twisted, i.e., surrounded in the negative sense, so that this part counts for negative area (Fig. 14.5). If time proceeded in the negative sense, *both* areas would increase, but the first area would be oriented negatively.

Between the two-dimensional invariant of Theorem 14.12 and the $2n$ -dimensional of Liouville's theorem, there are many others; e.g., the differential 4-form

$$\omega^4 = \sum_{i < j} dp_i \wedge dp_j \wedge dq_i \wedge dq_j. \quad (14.41)$$

These invariants, however, are not really new, because (14.41) is proportional to the exterior square of ω^2 , $\omega^2 \wedge \omega^2 = -2\omega^4$.

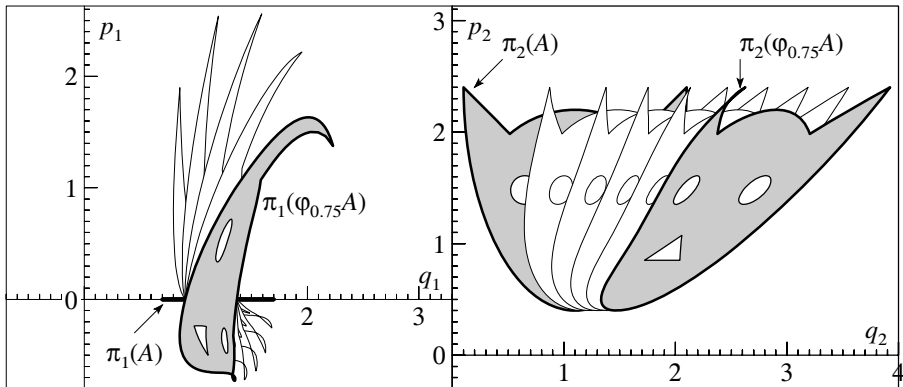


Fig. 14.5. Invariance of $\omega^2 = \sum_{i=1}^n dp_i \wedge dq_i$ for the spherical pendulum

Writing (14.31) in matrix notation

$$\omega^2(\xi_1, \xi_2) = \xi_1^T J \xi_2 \quad \text{with} \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \quad (14.42)$$

we obtain the following criterion:

Theorem 14.14. A differentiable transformation $g : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ is symplectic if and only if its Jacobian $R = g'(p, q)$ satisfies

$$R^T J R = J \quad (14.43)$$

with J given in (14.42).

Proof. This follows at once from (see (14.35))

$$(g^* \omega^2)(\xi_1, \xi_2) = (R\xi_1)^T J (R\xi_2) = \xi_1^T R^T J R \xi_2. \quad \square$$

Exercises

1. Prove the following lemma from elementary calculus which is used in the proof of Theorem 14.4: if for a function $F(x, y)$, $\partial F / \partial y$ exists and $y(x)$ is differentiable and such that $F(x, y(x)) = \text{Const}$, then $\partial F / \partial x$ exists at $(x, y(x))$ and is equal to

$$\frac{\partial F}{\partial x}(x, y(x)) = -\frac{\partial F}{\partial y}(x, y(x)) \cdot y'(x).$$

Hint. Use the identity

$$F(x_1, y(x_1)) - F(x_0, y(x_1)) = F(x_0, y(x_0)) - F(x_0, y(x_1)).$$

I.15 Boundary Value and Eigenvalue Problems

Although our book is mainly concerned with initial value problems, we want to include in this first chapter some properties of boundary and eigenvalue problems.

Boundary Value Problems

They arise in systems of differential equations, say

$$\begin{aligned}y_1' &= f_1(x, y_1, y_2), \\ y_2' &= f_2(x, y_1, y_2),\end{aligned}\tag{15.1}$$

when there is *no* initial point x_0 at which $y_1(x_0)$ and $y_2(x_0)$ are known simultaneously. Questions of existence and uniqueness then become much more complicated.

Example 1. Consider the differential equation

$$y'' = \exp(y) \quad \text{or} \quad y_1' = y_2, \quad y_2' = \exp(y_1) \tag{15.2a}$$

with the *boundary conditions*

$$y_1(0) = a, \quad y_1(1) = b. \tag{15.2b}$$

In order to apply our existence theorems or to do numerical computations (say by Euler's method (7.3)), we can proceed as follows: guess the missing initial value y_{20} . We can then compute the solution and check whether the computed value for $y_1(1)$ is equal to b or not. So our problem is, whether the function of the single variable y_{20}

$$F(y_{20}) := y_1(1) - b \tag{15.3}$$

possesses a zero or not.

Equation (15.2a) is *quasimonotone*, which implies that $F(y_{20})$ depends monotonically on y_{20} (Fig. 15.1a, see Exercise 7 of I.10). Also, for y_{20} very small or very large, $y_1(1)$ is arbitrarily small or large, or even infinite. Therefore, (15.2) possesses for all a, b a unique solution (see Fig. 15.1b).

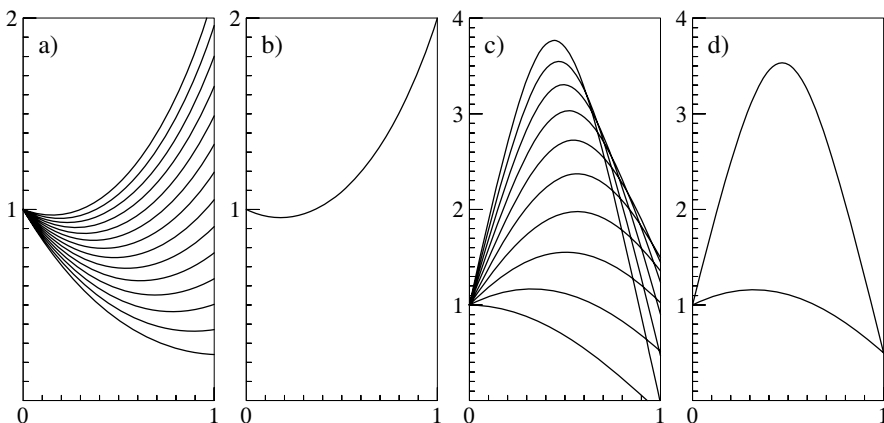


Fig. 15.1. a) Solutions of (15.2a) for different initial values $y_{20} = -1.7, \dots, -0.4$

b) Unique solution of (15.2a) for $a = 1, b = 2, y_{20} = -0.476984656$

c) Solutions of (15.4a) for $y(0) = 1$ and $y_{20} = 0, 1, 2, \dots, 9$

d) The two solutions of (15.4a), $y(0) = 1, y(1) = 0.5, y_{20} = 7.93719, y_{20} = 0.97084$

The root of $F(y_{20}) = 0$ can be computed by an iterative method, (bisection, regula falsi, . . . ; if the derivative of $y_1(1)$ with respect to y_{20} is used from Theorem 14.3 or numerically from finite differences, also by Newton's method). The initial value problem is then computed several times. Small problems, such as the above example, can be done by a simple dialogue with the computer. Harder problems with more unknown initial values need more programming skills. This method is one of the most commonly used and is called *the shooting method*.

Example 2. For the differential equation

$$y'' = -\exp(y) \quad \text{or} \quad y'_1 = y_2, \quad y'_2 = -\exp(y_1) \quad (15.4a)$$

with the *boundary conditions*

$$y_1(0) = a, \quad y_1(1) = b \quad (15.4b)$$

the monotonicity of $F(y_{20})$ is lost and things become more complicated: solutions for different initial values y_{20} are sketched for $a = 1$ in Fig. 15.1c. It can be seen that for b above a certain value (which is 1.499719998) there exists *no* solution of the problem at all, and for b below this value there exist *two* solutions (Fig. 15.1d).

Example 3.

$$y'_1 = y_2, \quad y'_2 = y_1^3, \quad y_1(0) = 1, \quad y_1(100) = 2. \quad (15.5)$$

This equation is similar to (15.2) and the same statement of existence and uniqueness holds as above. However, if one tries to compute the solutions by the shooting method, one gets into trouble because of the length of the interval: *the solution nearly never exists on the whole interval*; in fact, the correct solution is

$y_{20} = -0.70710616655$. But already for $y_{20} = -0.7071061$, $y_1(x)$ tends to $+\infty$ for $x \rightarrow 98.2$. On the other side, for $y_{20} = -0.70711$, we have $y_1(94.1) = -\infty$. So the domain where $F(y_{20})$ of (15.3) exists is of length less than 4×10^{-6} .

In a case like this, one can use the *multiple shooting technique*: the interval is split up into several subintervals, on each of which the problem is solved with well-chosen initial values. At the endpoints of the subintervals, the solutions are then matched together. Equation (15.3) thereby becomes a system of higher dimension to be solved. Another possibility is to apply *global methods* (finite differences, collocation). Instead of integrating a sequence of initial value problems, a global representation of the approximate solution is sought. There exists an extensive literature on methods for boundary value problems. As a general reference we give Ascher, Mattheij & Russel (1988) and Deuflhard (1980).

Sturm-Liouville Eigenvalue Problems

This subject originated with a remarkable paper of Sturm (Sturm 1836) in Liouville's newly founded Journal. This paper was followed by a series of papers by Liouville and Sturm published in the following volumes. It is today considered as the starting point of the “geometric theory”, where the main effort is not to try to integrate the equation, but merely to obtain geometric properties of the solution, such as its form, oscillations, sign changes, zeros, existence of maxima or minima and so on, *directly from the differential equation* (“Or on peut arriver à ce but par la seule considération des équations différentielles en elles-mêmes, sans qu'on ait besoin de leur intégration.”)

The physical origin was, as in Section I.6, the study of heat and small oscillations of elastic media. Let us consider the heat equation with non-constant conductivity

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left(k(x) \frac{\partial u}{\partial x} \right) - \ell(x)u, \quad k(x) > 0, \quad (15.6)$$

which was studied extensively in Poisson's “Théorie de la chaleur”. Poisson (1835) assumes $u(x, t) = y(x)e^{-\lambda t}$, so that (15.6) becomes

$$\frac{d}{dx} \left(k(x) \frac{dy}{dx} \right) - \ell(x)y = -\lambda y. \quad (15.7)$$

We write (15.7) in the form

$$(k(x)y')' + G(x)y = 0 \quad (15.8)$$

and state the following comparison theorem of Sturm:

Theorem 15.1. Consider, with (15.8), the differential equation

$$(\widehat{k}(x)\widehat{y}')' + \widehat{G}(x)\widehat{y} = 0, \quad (15.9)$$

and assume k, \widehat{k} differentiable, G, \widehat{G} continuous,

$$0 < \widehat{k}(x) \leq k(x), \quad \widehat{G}(x) \geq G(x) \quad (15.10)$$

for all x and let $y(x), \widehat{y}(x)$ be linearly independent solutions of (15.8) and (15.9), respectively. Then, between any two zeros of $y(x)$ there is at least one zero of $\widehat{y}(x)$, i.e., if $y(x_1) = y(x_2) = 0$ with $x_1 < x_2$ then there exists x_3 in the open interval (x_1, x_2) such that $\widehat{y}(x_3) = 0$.

Proof. The original proof of Sturm is based on the quotient

$$q(x) = \frac{y(x)}{k(x)y'(x)}$$

which is the slope of the line connecting the origin with the solution point in the (ky', y) -plane and satisfies a first-order differential equation. In order to avoid the singularities caused by the zeros of $y'(x)$, we prefer the use of polar coordinates (Prüfer 1926)

$$k(x)y'(x) = \varrho(x) \cos \varphi(x), \quad y(x) = \varrho(x) \sin \varphi(x). \quad (15.11)$$

Differentiation of (15.11) yields the following differential equations for φ and ϱ :

$$\varphi' = \frac{1}{k(x)} \cos^2 \varphi + G(x) \sin^2 \varphi \quad (15.12)$$

$$\varrho' = \left(\frac{1}{k(x)} - G(x) \right) \cdot \sin \varphi \cdot \cos \varphi \cdot \varrho. \quad (15.13)$$

In the same way we also introduce functions $\widehat{\varrho}(x)$ and $\widehat{\varphi}(x)$ for the second differential equation (15.9). They satisfy analogous relations with $k(x)$ and $G(x)$ replaced by $\widehat{k}(x)$ and $\widehat{G}(x)$.

Suppose now that x_1, x_2 are two consecutive zeros of $y(x)$. Then $\varphi(x_1)$ and $\varphi(x_2)$ must be multiples of π , since $\varrho(x)$ is always different from zero (uniqueness of the initial value problem). By (15.12) $\varphi'(x)$ is positive at x_1 and at x_2 . Therefore we may assume that

$$\varphi(x_1) = 0, \quad \varphi(x_2) = \pi, \quad \widehat{\varphi}(x_1) \in [0, \pi). \quad (15.14)$$

The fact that equation (15.12) is first-order and the inequalities (15.10) allow the application of Theorem 10.3 to give

$$\widehat{\varphi}(x) \geq \varphi(x) \quad \text{for} \quad x_1 \leq x \leq x_2.$$

It is impossible that $\widehat{\varphi}(x) = \varphi(x)$ everywhere, since this would imply $\widehat{G}(x) = G(x)$, $\cos \widehat{\varphi}(x)/\widehat{k}(x) = \cos \varphi(x)/k(x)$ by (15.12) and (15.10). As a consequence of (15.13) we would have $\widehat{\varrho}(x) = C \cdot \varrho(x)$ and the solutions $y(x), \widehat{y}(x)$ would be

linearly dependent. Therefore, there exists $x_0 \in (x_1, x_2)$ such that $\widehat{\varphi}(x_0) > \varphi(x_0)$. In this situation $\widehat{\varphi}(x) > \varphi(x)$ for all $x \geq x_0$ and the existence of $x_3 \in (x_1, x_2)$ with $\widehat{\varphi}(x_3) = \pi$ is assured. \square

The next theorem shows that our eigenvalue problem possesses an *infinity* of solutions. We add to (15.7) the boundary conditions

$$y(x_0) = y(x_1) = 0. \quad (15.15)$$

Theorem 15.2. *The eigenvalue problem (15.7), (15.15) possesses an infinite sequence of eigenvalues $\lambda_1 < \lambda_2 < \lambda_3 < \dots$ whose corresponding solutions $y_i(x)$ (“eigenfunctions”) possess respectively $0, 1, 2, \dots$ zeros in the interval (x_0, x_1) . The zeros of $y_{j+1}(x)$ separate those of $y_j(x)$. If $0 < K_1 \leq k(x) \leq K_2$ and $L_1 \leq \ell(x) \leq L_2$, then*

$$L_1 + K_1 \frac{j^2 \pi^2}{(x_1 - x_0)^2} \leq \lambda_j \leq L_2 + K_2 \frac{j^2 \pi^2}{(x_1 - x_0)^2}. \quad (15.16)$$

Proof. Let $y(x, \lambda)$ be the solution of (15.7) with initial values $y(x_0) = 0$, $y'(x_0) = 1$. Theorem 15.1 (with $\widehat{k}(x) = k(x)$, $\widehat{G}(x) = G(x) + \Delta\lambda$) implies that for increasing λ the zeros of $y(x, \lambda)$ move towards x_0 , so that the number of zeros in (x_0, x_1) is a non-decreasing function of λ .

Comparing next (15.7) with the solution ($\lambda > L_1$)

$$\sin\left(\sqrt{(\lambda - L_1)/K_1} \cdot (x - x_0)\right)$$

of $K_1 y'' + (\lambda - L_1)y = 0$ we see that for $\lambda < L_1 + K_1 j^2 \pi^2 / (x_1 - x_0)^2$, $y(x, \lambda)$ has at most $j - 1$ zeros in $(x_0, x_1]$. Similarly, a comparison with

$$\sin\left(\sqrt{(\lambda - L_2)/K_2} \cdot (x - x_0)\right)$$

which is a solution of $K_2 y'' + (\lambda - L_2)y = 0$, shows that $y(x, \lambda)$ possesses at least j zeros in (x_0, x_1) , if $\lambda > L_2 + K_2 j^2 \pi^2 / (x_1 - x_0)^2$. The statements of the theorem are now simple consequences of these three properties. \square

Example. Fig. 15.2 shows the first 5 solutions of the problem

$$((1 - 0.8 \sin^2 x)y')' - (x - \lambda)y = 0, \quad y(0) = y(\pi) = 0. \quad (15.17)$$

The first eigenvalues are 2.1224, 3.6078, 6.0016, 9.3773, 13.7298, 19.053, 25.347, 32.609, 40.841, 50.041, etc.

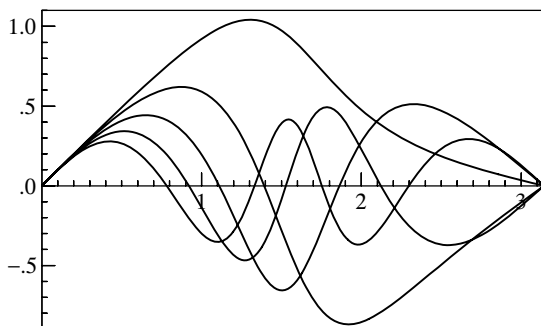


Fig. 15.2. Solutions of the Sturm-Liouville eigenvalue problem (15.17)

For more details about this theory, which is a very important page of history, we refer to the book of Reid (1980).

Exercises

1. Consider the equation

$$L(x)y'' + M(x)y' + N(x)y = 0.$$

Multiply it with a suitable function $\varphi(x)$, so that the ensuing equation is of the form (15.8) (Sturm 1836, p. 108).

2. Prove that two solutions of (15.7), (15.15) satisfy the orthogonality relations

$$\int_{x_0}^{x_1} y_j(x)y_k(x)dx = 0 \quad \text{for} \quad \lambda_j \neq \lambda_k.$$

Hint. Multiply this by λ_j , replace $\lambda_j y_j(x)$ from (15.7) and do partial integration (Liouville 1836, p. 257).

3. Solve the problem (15.5) by elementary functions. Explain why the given value for y_{20} is so close to $-\sqrt{2}/2$.
4. Show that the boundary value problem (see Collatz 1967)

$$y'' = -y^3, \quad y(0) = 0, \quad y(A) = B \quad (15.18)$$

possesses infinitely many solutions for each pair (A, B) with $A \neq 0$.

Hint. Draw the solution $y(x)$ of (15.18) with $y(0) = 0$, $y'(0) = 1$. Show that for each constant a , $z(x) = ay(ax)$ is also a solution.

I.16 Periodic Solutions, Limit Cycles, Strange Attractors

2° Les demi-spirales que l'on suit sur un arc infini sans arriver à un nœud ou à un foyer et sans revenir au point de départ ; ...
(H. Poincaré 1882, Oeuvres vol. 1, p. 54)

The phenomenon of limit cycles was first described theoretically by Poincaré (1882) and Bendixson (1901), and has since then found many applications in Physics, Chemistry and Biology. In higher dimensions things can become much more chaotic and attractors may look fairly “strange”.

Van der Pol's Equation

I have a theory that whenever you want to get in trouble with a method, look for the Van der Pol equation.
(P.E. Zadunaisky 1982)

The first practical examples were studied by Rayleigh (1883) and later by Van der Pol (1920-1926) in a series of papers on nonlinear oscillations: the solutions of

$$y'' + \alpha y' + y = 0$$

are *damped* for $\alpha > 0$, and *unstable* for $\alpha < 0$. The idea is to change α (with the help of a triode, for example) so that $\alpha < 0$ for small y and $\alpha > 0$ for large y . The simplest expression, which describes the physical situation in a somewhat idealized form, would be $\alpha = \varepsilon(y^2 - 1)$, $\varepsilon > 0$. Then the above equation becomes

$$y'' + \varepsilon(y^2 - 1)y' + y = 0, \quad (16.1)$$

or, written as a system,

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= \varepsilon(1 - y_1^2)y_2 - y_1, \quad \varepsilon > 0. \end{aligned} \quad (16.2)$$

In this equation, small oscillations are amplified and large oscillations are damped. We therefore expect the existence of a stable periodic solution to which all other solutions converge. We call this a *limit cycle* (Poincaré 1882, “Chap. VI”). The original illustrations of the paper of Van der Pol are reproduced in Fig. 16.1.

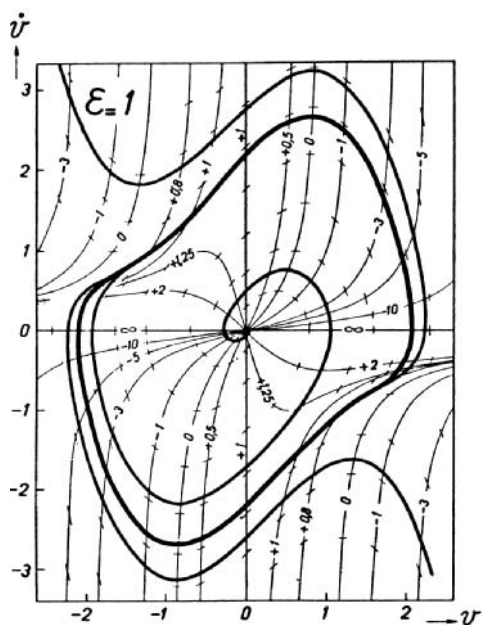
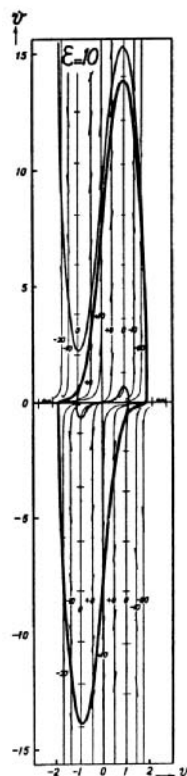


Fig. 16.1. Illustrations from
Van der Pol (1926)
(with permission)



Existence proof. The existence of limit cycles is studied by the method of *Poincaré sections* (Poincaré 1882, “Chap. V, Théorie des conséquents”). The idea is to cut the solutions transversally by a hyperplane Π and, for an initial value $y_0 \in \Pi$, to study the first point $\Phi(y_0)$ where the solution again crosses the plane Π in the same direction.

For our example (16.2), we choose for Π the half-line $y_2 = 0, y_1 > 0$. We then examine the signs of y_1' and y_2' in (16.2). The sign of y_2' changes at the curve

$$y_2 = \frac{y_1}{\varepsilon(1 - y_1^2)}, \quad (16.3)$$

which is drawn as a broken line in Fig. 16.2. It follows (see Fig. 16.2) that $\Phi(y_0)$ exists for all $y_0 \in \Pi$. Since two different solutions *cannot intersect* (due to uniqueness), the map Φ is *monotone*. Further, Φ is bounded (e.g., by every solution starting on the curve (16.3)), so $\Phi(y_0) < y_0$ for y_0 large. Finally, since the origin is unstable, $\Phi(y_0) > y_0$ for y_0 small. Hence there must be a fixed point of $\Phi(y_0)$, i.e., a limit cycle. \square

The limit cycle is, in fact, *unique*. The proof for this is more complicated and is indicated in Exercise 8 below (Liénard 1928).

solved to give

$$\log(y_1) - \frac{y_1^2}{2} = \frac{x - x_0}{\varepsilon} + \text{Const.}$$

These curves are dotted in Van der Pol's Fig. 16.3 for $\varepsilon = 10$ and show the good approximation of this solution.

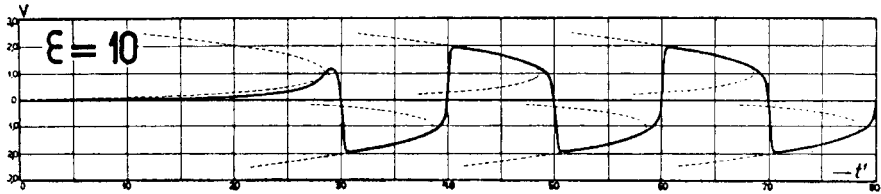


Fig. 16.3. Solution of Van der Pol's equation for $\varepsilon = 10$ compared with steady state approximations

Asymptotic solutions for ε small. The computation of periodic solutions for *small* parameters was initiated by astronomers such as Newcomb and Lindstedt and brought to perfection by Poincaré (1893). We demonstrate the method for the Van der Pol equation (16.1). The idea is to develop the solution as a series in powers of ε . Since the period will change too, we also introduce a coordinate change

$$t = x(1 + \gamma_1\varepsilon + \gamma_2\varepsilon^2 + \dots) \quad (16.6)$$

and put

$$y(x) = z(t) = z_0(t) + \varepsilon z_1(t) + \varepsilon^2 z_2(t) + \dots \quad (16.7)$$

Inserting now $y'(x) = z'(t)(1 + \gamma_1\varepsilon + \dots)$, $y''(x) = z''(t)(1 + \gamma_1\varepsilon + \dots)^2$ into (16.1) we obtain

$$\begin{aligned} & (z_0'' + \varepsilon z_1'' + \varepsilon^2 z_2'' + \dots)(1 + 2\gamma_1\varepsilon + (2\gamma_2 + \gamma_1^2)\varepsilon^2 + \dots) \\ & + \varepsilon((z_0 + \varepsilon z_1 + \dots)^2 - 1)(z_0' + \varepsilon z_1' + \dots)(1 + \gamma_1\varepsilon + \dots) \\ & + (z_0 + \varepsilon z_1 + \varepsilon^2 z_2 + \dots) = 0. \end{aligned} \quad (16.8)$$

We first compare the coefficients of ε^0 and obtain

$$z_0'' + z_0 = 0. \quad (16.8;0)$$

We fix the initial value on the Poincaré section P , i.e., $z'(0) = 0$, so that $z_0 = A \cos t$ with A , for the moment, a free parameter. Next, the coefficients of ε yield

$$\begin{aligned} z_1'' + z_1 &= -2\gamma_1 z_0'' - (z_0^2 - 1)z_0' \\ &= 2\gamma_1 A \cos t + \left(\frac{A^3}{4} - A\right) \sin t + \frac{A^3}{4} \sin 3t. \end{aligned} \quad (16.8;1)$$

Here, the crucial idea is that we are looking for *periodic* solutions, hence the terms in $\cos t$ and $\sin t$ on the right-hand side of (16.8;1) must disappear, in order to avoid that $z_1(t)$ contain terms of the form $t \cdot \cos t$ and $t \cdot \sin t$ (“... et de faire disparaître ainsi les termes dits *séculaires* ...”). We thus obtain $\gamma_1 = 0$ and $A = 2$. Then (16.8;1) can be solved and gives, together with $z_1'(0) = 0$,

$$z_1 = B \cos t + \frac{3}{4} \sin t - \frac{1}{4} \sin 3t. \quad (16.9)$$

The continuation of this process is now clear: the terms in ε^2 in (16.8) lead to, after insertion of (16.9) and simplification,

$$z_2'' + z_2 = \left(4\gamma_2 + \frac{1}{4}\right) \cos t + 2B \sin t + 3B \sin 3t - \frac{3}{2} \cos 3t + \frac{5}{4} \cos 5t. \quad (16.8;2)$$

Secular terms are avoided if we set $B = 0$ and $\gamma_2 = -1/16$. Then

$$z_2 = C \cos t + \frac{3}{16} \cos 3t - \frac{5}{96} \cos 5t.$$

The next round will give $C = -1/8$ and $\gamma_3 = 0$, so that we have: *the periodic orbit of the Van der Pol equation (16.1) for ε small is given by*

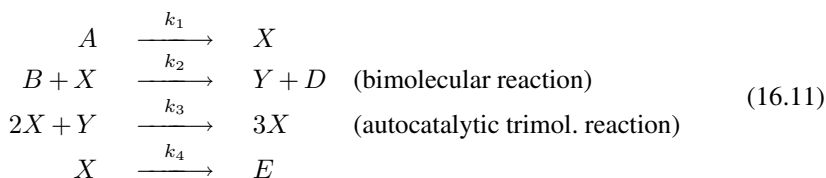
$$\begin{aligned} y(x) &= z(t), & t &= x(1 - \varepsilon^2/16 + \dots), \\ z(t) &= 2 \cos t + \varepsilon \left(\frac{3}{4} \sin t - \frac{1}{4} \sin 3t \right) \\ &\quad + \varepsilon^2 \left(-\frac{1}{8} \cos t + \frac{3}{16} \cos 3t - \frac{5}{96} \cos 5t \right) + \dots \end{aligned} \quad (16.10)$$

and is of period $2\pi(1 + \varepsilon^2/16 + \dots)$.

Chemical Reactions

The laws of chemical kinetics give rise to differential equations which, for multi-molecular reactions, become nonlinear and have interesting properties. Some of them possess periodic solutions (e.g. the Zhabotinski-Belousov reaction) and have important applications to the interpretation of biological phenomena (e.g. Prigogine, Lefever).

Let us examine in detail the model of Lefever and Nicolis (1971), the so-called “Brusselator”: suppose that six substances A, B, D, E, X, Y undergo the following reactions:



If we denote by $A(x), B(x), \dots$ the *concentrations* of A, B, \dots as functions of the time x , the reactions (16.11) become by the mass action law the following differential equations

$$\begin{aligned} A' &= -k_1 A \\ B' &= -k_2 B X \\ D' &= k_2 B X \\ E' &= k_4 X \\ X' &= k_1 A - k_2 B X + k_3 X^2 Y - k_4 X \\ Y' &= k_2 B X - k_3 X^2 Y. \end{aligned}$$

This system is now simplified as follows: the equations for D and E are left out, because they do not influence the others; A and B are supposed to be maintained constant (positive) and all reaction rates k_i are set equal to 1. We further set $y_1(x) := X(x)$, $y_2(x) := Y(x)$ and obtain

$$\begin{aligned} y_1' &= A + y_1^2 y_2 - (B + 1)y_1 \\ y_2' &= B y_1 - y_1^2 y_2. \end{aligned} \tag{16.12}$$

The resulting system has one critical point $y_1' = y_2' = 0$ at $y_1 = A$, $y_2 = B/A$. The linearized equation in the neighbourhood of this point is unstable iff $B > A^2 + 1$. Further, a study of the domains where y_1' , y_2' , or $(y_1 + y_2)'$ is positive or negative leads to the result that all solutions remain bounded. Thus, for $B > A^2 + 1$ there must be a limit cycle which, by numerical calculations, is seen to be unique (Fig. 16.4).

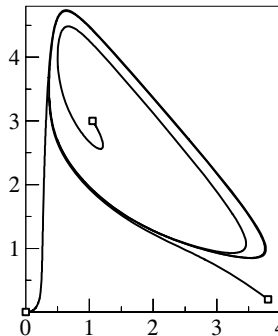


Fig. 16.4. Solutions of the Brusselator, $A = 1$, $B = 3$

An interesting phenomenon (Hopf bifurcation, see below) occurs, when B approaches $A^2 + 1$. Then the limit cycle becomes smaller and smaller and finally disappears in the critical point. Another example of this type is given in Exercise 2.

Limit Cycles in Higher Dimensions, Hopf Bifurcation

The Theorem of Poincaré-Bendixson is apparently true only in two dimensions. Higher dimensional counter-examples are given by nearly every mechanical movement without friction, as for example the spherical pendulum (6.20), see Fig. 6.2. Therefore, in higher dimensions limit cycles are usually found by numerical studies of the Poincaré section map Φ defined above.

There is, however, one situation where limit cycles occur quite naturally (Hopf 1942): namely when at a critical point of $y' = f(y, \alpha)$, $y, f \in \mathbb{R}^n$, all eigenvalues of $(\partial f / \partial y)(y_0, \alpha)$ have strictly negative real part with the exception of *one* pair which, by varying α , crosses the imaginary axis. The eigenspace of the stable eigenvalues then continues into an analytic two dimensional manifold, inside which a limit cycle appears. This phenomenon is called “Hopf bifurcation”. The proof of this fact is similar to Poincaré’s parameter expansion method (16.7) (see Exercises 6 and 7 below), so that Hopf even hesitated to publish it (“... ich glaube kaum, dass an dem obigen Satz etwas wesentlich Neues ist ...”).

As an example, we consider the “full Brusselator” (16.11): we no longer suppose that B is kept constant, but that B is constantly added to the mixture with

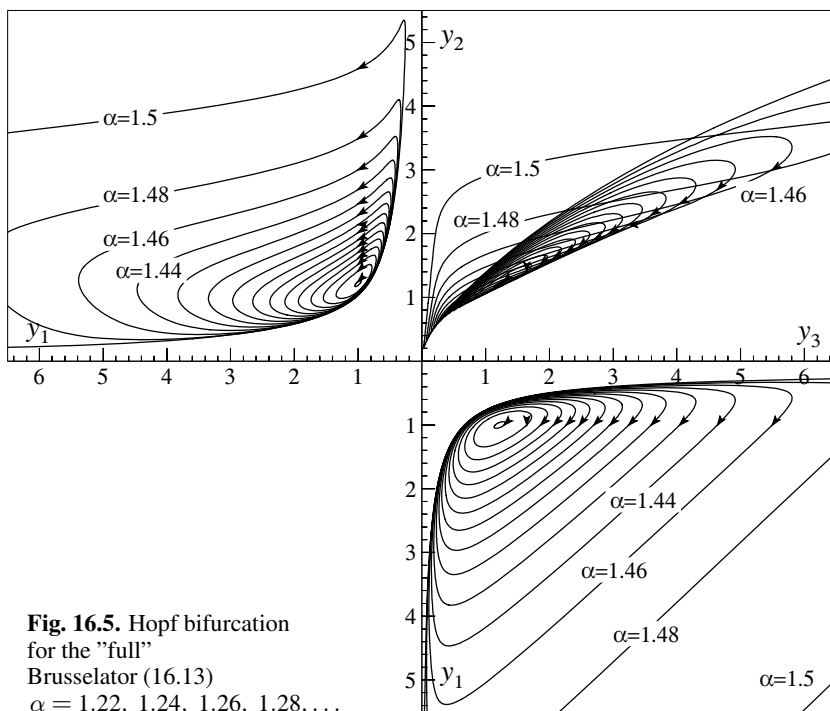


Fig. 16.5. Hopf bifurcation for the “full” Brusselator (16.13)
 $\alpha = 1.22, 1.24, 1.26, 1.28, \dots$

rate α . When we set $y_3(x) := B(x)$, we obtain instead of (16.12) (with $A = 1$)

$$\begin{aligned} y_1' &= 1 + y_1^2 y_2 - (y_3 + 1)y_1 \\ y_2' &= y_1 y_3 - y_1^2 y_2 \\ y_3' &= -y_1 y_3 + \alpha. \end{aligned} \quad (16.13)$$

This system possesses a critical point at $y_1 = 1$, $y_2 = y_3 = \alpha$ with derivative

$$\frac{\partial f}{\partial y} = \begin{pmatrix} \alpha - 1 & 1 & -1 \\ -\alpha & -1 & 1 \\ -\alpha & 0 & -1 \end{pmatrix}. \quad (16.14)$$

This matrix has $\lambda^3 + (3 - \alpha)\lambda^2 + (3 - 2\alpha)\lambda + 1$ as characteristic polynomial and satisfies the condition for stability iff $\alpha < (9 - \sqrt{17})/4 = 1.21922$ (see I.13, Exercise 1). Thus when α increases beyond this value, there arises a limit cycle which exists for all values of α up to approximately 1.5 (see Fig. 16.5). When α continues to grow, the limit cycle “explodes” and $y_1 \rightarrow 0$ while y_2 and $y_3 \rightarrow \infty$. So the system (16.13) has a behaviour completely different from the simplified model (16.12).

A famous chemical reaction with a limit cycle in three dimensions is the “Oregonator” reaction between $HBrO_2$, Br^- , and $Ce(IV)$ (Field & Noyes 1974)

$$\begin{aligned} y_1' &= 77.27 \left(y_2 + y_1 (1 - 8.375 \times 10^{-6} y_1 - y_2) \right) \\ y_2' &= \frac{1}{77.27} (y_3 - (1 + y_1)y_2) \\ y_3' &= 0.161 (y_1 - y_3) \end{aligned} \quad (16.15)$$

whose solutions are plotted in Fig. 16.6. This is an example of a “stiff” differential equation whose solutions change rapidly over many orders of magnitude. It is thus a challenging example for numerical codes and we shall meet it again in Volume II of our book.

Our next example is taken from the theory of superconducting Josephson junctions, coupled together by a mutual capacitance. Omitting all physical details, (see Giovannini, Weiss & Ulrich 1978), we state the resulting equations as

$$\begin{aligned} c(y_1'' - \alpha y_2'') &= i_1 - \sin(y_1) - y_1' \\ c(y_2'' - \alpha y_1'') &= i_2 - \sin(y_2) - y_2'. \end{aligned} \quad (16.16)$$

Here, y_1 and y_2 are *angles* (the “quantum phase difference across the junction”) which are thus identified modulo 2π . Equation (16.16) is thus a system on the torus T^2 for (y_1, y_2) , and on \mathbb{R}^2 for the voltages (y_1', y_2') . It is seen by numerical computations that the system (16.16) possesses an attracting limit cycle, which describes the phenomenon of “phase locking” (see Fig. 16.7).

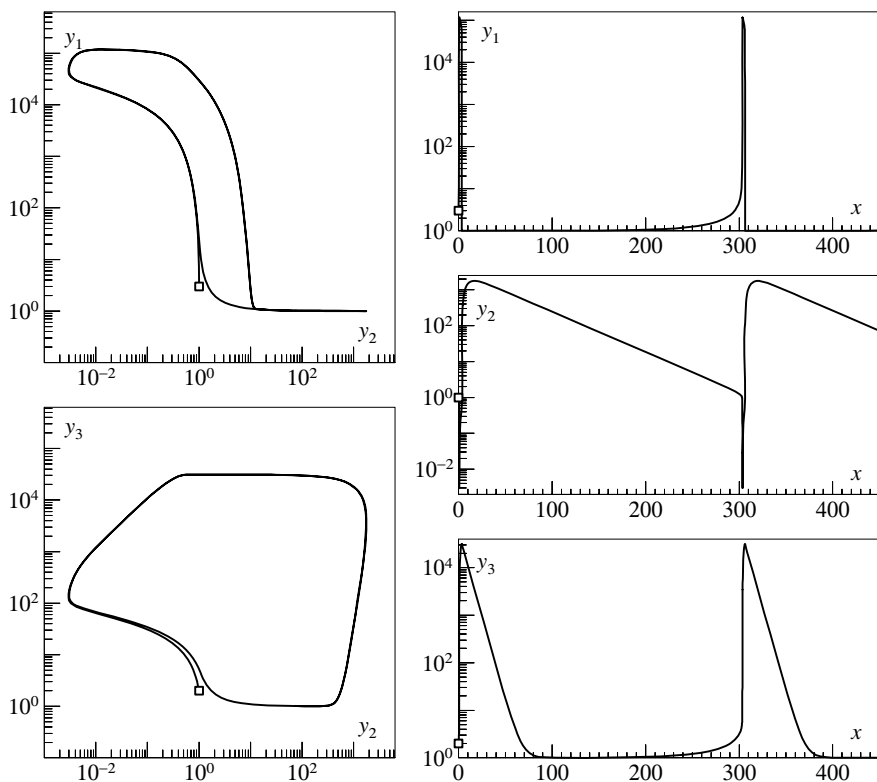


Fig. 16.6. Limit cycle of the Oregonator

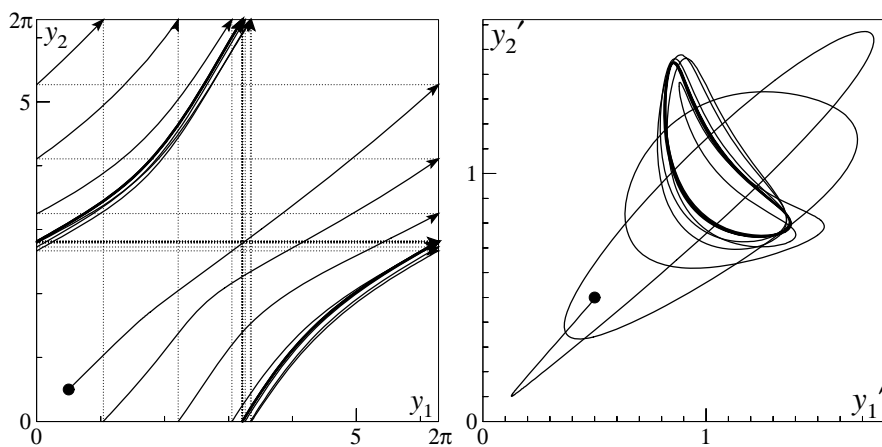


Fig. 16.7. Josephson junctions (16.16) for $c = 2$, $\alpha = 0.5$, $i_1 = 1.11$, $i_2 = 1.08$

Strange Attractors

“Mr. Dahlquist, when is the spring coming ?”

“Tomorrow, at two o’clock.”

(Weather forecast, Stockholm 1955)

“We were **so** naïve . . .”

(H.O. Kreiss, Stockholm 1985)

Concerning the discovery of the famous “Lorenz model”, we best quote from Lorenz (1979):

“By the middle 1950’s “numerical weather prediction”, i.e., forecasting by numerically integrating such approximations to the atmospheric equations as could feasibly be handled, was very much in vogue, despite the rather mediocre results which it was then yielding. A smaller but determined group favored statistical prediction (. . .) apparently because of a misinterpretation of a paper by Wiener (. . .). I was skeptical, and decided to test the idea by applying the statistical method to a set of artificial data, generated by solving a system of equations numerically (. . .). The first task was to find a suitable system of equations to solve (. . .). The system would have to be simple enough (. . . and) the general solution would have to be aperiodic, since the statistical prediction of a periodic series would be a trivial matter, once the periodicity had been detected. It was not obvious that these conditions could be met. (. . .) The break came when I was visiting Dr. Barry Saltzman, now at Yale University. In the course of our talks he showed me some work on thermal convection, in which he used a system of seven ordinary differential equations. Most of his numerical solutions soon acquired periodic behavior, but one solution refused to settle down. Moreover, in this solution four of the variables appeared to approach zero. Presumably the equations governing the remaining three variables, with the terms containing the four variables eliminated, would also possess aperiodic solutions. Upon my return I put the three equations on our computer, and confirmed the aperiodicity which Saltzman had noted. We were finally in business.”

In a changed notation, the three equations with aperiodic solutions are

$$\begin{aligned}y_1' &= -\sigma y_1 + \sigma y_2 \\y_2' &= -y_1 y_3 + r y_1 - y_2 \\y_3' &= y_1 y_2 - b y_3\end{aligned}\tag{16.17}$$

where σ , r and b are positive constants. It follows from (16.17) that

$$\begin{aligned}\frac{1}{2} \frac{d}{dx} \left(y_1^2 + y_2^2 + (y_3 - \sigma - r)^2 \right) \\= - \left(\sigma y_1^2 + y_2^2 + b \left(y_3 - \frac{\sigma}{2} - \frac{r}{2} \right)^2 \right) + b \left(\frac{\sigma}{2} + \frac{r}{2} \right)^2.\end{aligned}\tag{16.18}$$

Therefore the ball

$$R_0 = \left\{ (y_1, y_2, y_3) \mid y_1^2 + y_2^2 + (y_3 - \sigma - r)^2 \leq c^2 \right\} \quad (16.19)$$

is mapped by the flow φ_1 (see (14.22)) into itself, provided that c is sufficiently large so that R_0 wholly contains the ellipsoid defined by equating the right side of (16.18) to zero. Hence, if x assumes the increasing values $1, 2, 3, \dots$, R_0 is carried into regions $R_1 = \varphi_1(R_0)$, $R_2 = \varphi_2(R_0)$ etc., which satisfy $R_0 \supset R_1 \supset R_2 \supset R_3 \supset \dots$ (applying φ_1 to the inclusion $R_0 \supset R_1$ gives $R_1 \supset R_2$ and so on).

Since the trace of $\partial f / \partial y$ for the system (16.17) is the negative constant $-(\sigma + b + 1)$, the *volumes* of R_k tend exponentially to zero (see Theorem 14.8). Every orbit is thus ultimately trapped in a set $R_\infty = R_0 \cap R_1 \cap R_2 \dots$ of zero volume.

System (16.17) possesses an obvious critical point $y_1 = y_2 = y_3 = 0$; this becomes unstable when $r > 1$. In this case there are two additional critical points C and C' respectively given by

$$y_1 = y_2 = \pm \sqrt{b(r-1)}, \quad y_3 = r - 1. \quad (16.20)$$

These become unstable (e.g. by the Routh criterion, Exercise 1 of Section I.13) when $\sigma > b + 1$ and

$$r \geq r_c = \frac{\sigma(\sigma + b + 3)}{\sigma - b - 1}. \quad (16.21)$$

In the first example we shall use Saltzman's values $b = 8/3$, $\sigma = 10$, and $r = 28$. ("Here we note another lucky break: Saltzman used $\sigma = 10$ as a crude approximation to the Prandtl number (about 6) for water. Had he chosen to study air, he would probably have let $\sigma = 1$, and the aperiodicity would not have been discovered", Lorenz 1979). In Fig. 16.8 we have plotted the solution curve of (16.17) with the initial value $y_1 = -8$, $y_2 = 8$, $y_3 = r - 1$, which, indeed, looks pretty chaotic.

For a clearer understanding of the phenomenon, we choose the plane $y_3 = r - 1$, especially the square region between the critical points C and C' , as Poincaré section Π . The critical point $y_1 = y_2 = y_3 = 0$ possesses (since $r > 1$) one unstable eigenvalue $\lambda_1 = (-1 - \sigma + \sqrt{(1 - \sigma)^2 + 4r\sigma})/2$ and two stable eigenvalues $\lambda_2 = -b$, $\lambda_3 = (-1 - \sigma - \sqrt{(1 - \sigma)^2 + 4r\sigma})/2$. The eigenspace of the stable eigenvalues continues into a two-dimensional manifold of initial values, whose solutions tend to 0 for $x \rightarrow \infty$. This "stable manifold" cuts Π in a curve Σ (see Fig. 16.9). The one-dimensional *unstable* manifold (created by the unstable eigenvalue λ_1) cuts Π in the points D and D' (Fig. 16.9).

All solutions starting in Π_u *above* Σ (the dark cat) surround the above critical point C and are, at the first return, mapped to a narrow stripe S_u , while the solutions starting in Π_d *below* Σ surround C' and go to the left stripe S_d . At the *second* return, the two stripes are mapped into two very narrow stripes *inside* S_u and S_d . After the third return, we have 8 stripes closer and closer together, and so on. The intersection of all these stripes is a Cantor-like set and, continued

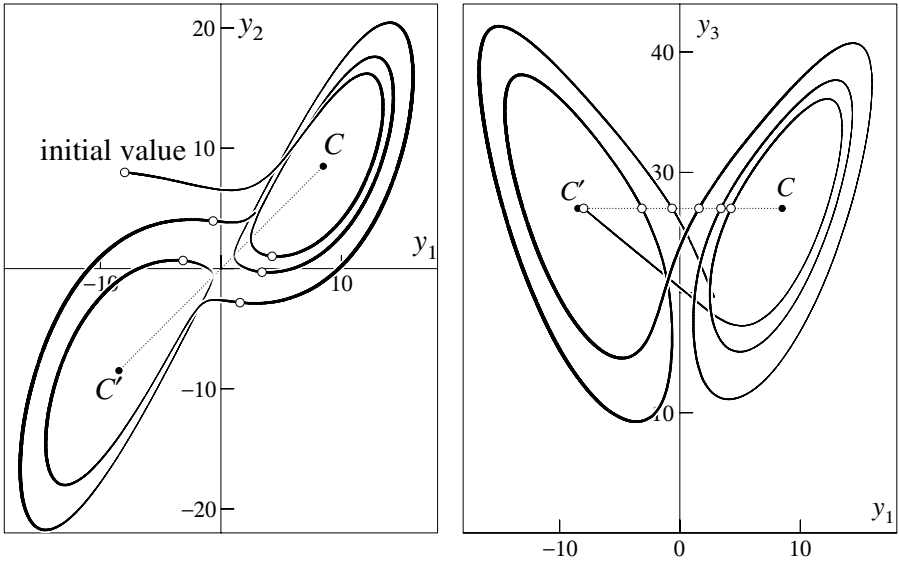


Fig. 16.8. Two views of a solution of (16.17)
(small circles indicate intersection of solution with plane $y_3 = r - 1$)

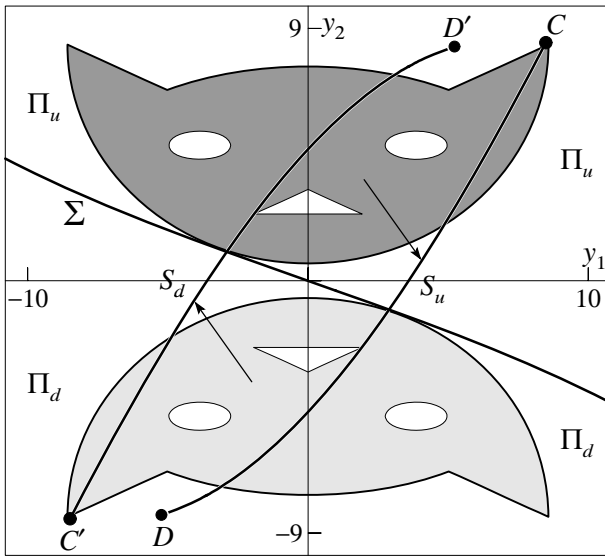
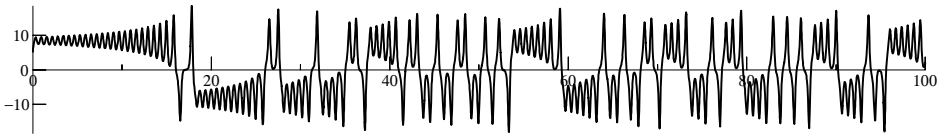


Fig. 16.9. Poincaré map for (16.17)

into 3-space by the flow, forms the *strange attractor* (“An attractor of the type just described can therefore not be thrown away as non-generic pathology”, Ruelle & Takens 1971).

The Ups and Downs of the Lorenz Model

“Mr. Laurel and Mr. Hardy have many ups and downs — Mr. Hardy takes charge of the upping, and Mr. Laurel does most of the downing —”
(from “Another Fine Mess”, Hal Roach 1930)



If one watches the solution $y_1(x)$ of the Lorenz equation being calculated, one wonders who decides for the solution to go up or down in an apparently unpredictable fashion. Fig. 16.9 shows that Σ cuts both stripes S_d and S_u . Therefore the *inverse image* of Σ (see Fig. 16.10) consists of *two* lines Σ_0 and Σ_1 which cut, together with Σ , the plane Π into *four* sets Π_{uu} , Π_{ud} , Π_{du} , Π_{dd} . If the initial value is in one of these, the corresponding solution goes up-up, up-down, down-up, down-down. Further, the inverse images of Σ_0 and Σ_1 lead to four lines Σ_{00} , Σ_{01} , Σ_{10} , Σ_{11} . The plane Π is then cut into 8 stripes and we now know the fate of the first three ups and downs. The more inverse images of these curves we compute, the finer the plane Π is cut into stripes and all the future ups and downs are coded in the position of the initial value with respect to these stripes (see Fig. 16.10). It appears that a *very small* change in the initial value gives rise, after a couple of rotations, to a *totally different* solution curve. This phenomenon, discovered merely by accident by Lorenz (see Lorenz 1979), is highly interesting

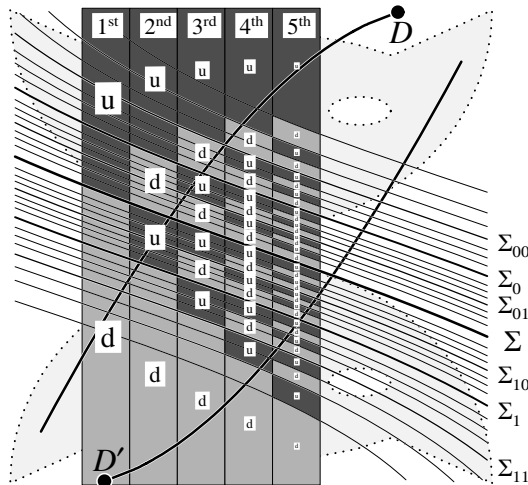


Fig. 16.10. Stripes deciding for the ups and downs

and explains why the theorem of uniqueness (Theorem 7.4), of whose philosophical consequences Laplace was so proud, has its practical limits.

Remark. It appears in Fig. 16.10 that not all stripes have the same width. The sequences of “*u*”s and “*d*”s which repeat *u* or *d* a couple of times (but not too often) are more probable than the others. More than 25 consecutive “ups” or “downs” are (for the chosen constants and except for the initial phase) never possible. This has to do with the position of *D* and *D'*, the outermost frontiers of the attractor, in the stripes of Fig. 16.10.

Feigenbaum Cascades

However nicely the beginning of Lorenz’ (1979) paper is written, the affirmations of his last section are only partly true. As Lorenz did, we now vary the parameter *b* in (16.17), letting at the same time $r = r_c$ (see (16.21)) and

$$\sigma = b + 1 + \sqrt{2(b+1)(b+2)}. \quad (16.22)$$

This is the value of σ for which r_c is minimized. Numerical integration shows that for *b* very small (say $b \leq 0.139$), the solutions of (16.17) evidently converge to a stable limit cycle, which cuts the Poincaré section $y_3 = r - 1$ twice at two different locations and surrounds both critical points *C* and *C'*. Further, for *b* large (for example $b = 8/3$) the coefficients are not far from those studied above and we have a strange attractor. But what happens in between? We have computed the solutions of the Lorenz model (16.17) for *b* varying from 0.1385 to 0.1475 with 1530 intermediate values. For each of these values, we have computed 1500 Poincaré cuts and represented in Fig. 16.11 the y_1 -values of the intersections with the Poincaré plane $y_3 = r - 1$. After each change of *b*, the first 300 iterations were not drawn so that only the attractor becomes visible.

For *b* small, there is one periodic orbit; then, at $b = b_1 = 0.13972$, it suddenly splits into an orbit of period two, this then splits for $b = b_2 = 0.14327$ into an orbit of period four, then for $b = b_3 = 0.14400$ into period eight, etc. There is a point $b_\infty = 0.14422$ after which the movement becomes chaotic. Beyond this value, however, there are again and again intervals of stable attractors of periods 5, 3, etc. The whole picture resembles what is obtained by the recursion

$$x_{n+1} = a(x_n - x_n^2) \quad (16.23)$$

which is discussed in many papers (e.g. May 1976, Feigenbaum 1978, Collet & Eckmann 1980).

But where does this resemblance come from? We study in Fig. 16.12 the Poincaré map for the system (16.17) with *b* chosen as 0.146 of a region $-0.095 \leq y_1 \leq -0.078$ and $-0.087 \leq y_2 \leq -0.07$. After one return, this region is compressed to a thin line somewhere else on the plane (Fig. 16.12b), the second return bends this line to *U*-shape and maps it into the original region (Fig. 16.12c).

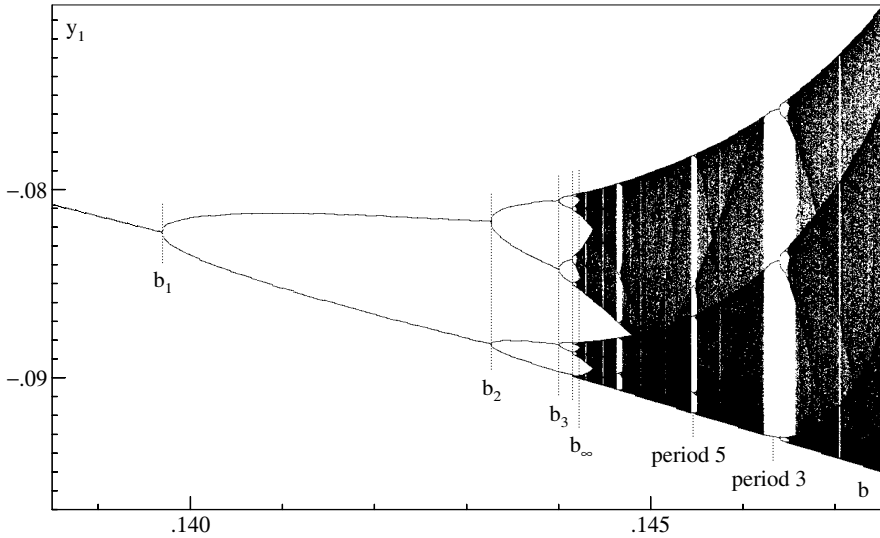


Fig. 16.11. Poincaré cuts y_1 for (16.17) as function of b

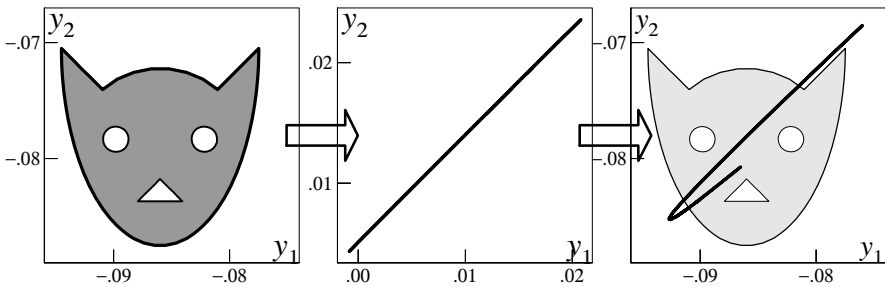


Fig. 16.12. Poincaré map for system (16.17) with $b = 0.146$

Therefore, the Poincaré map is essentially a map of the interval $[0, 1]$ to itself similar to (16.23). It is a great discovery of Feigenbaum that for *all* maps of a similar shape, the phenomena are always the same, in particular that

$$\lim_{i \rightarrow \infty} \frac{b_i - b_{i-1}}{b_{i+1} - b_i} = 4.6692016091029906715 \dots$$

is a universal constant, the *Feigenbaum number*. The repeated doublings of the periods at b_1, b_2, b_3, \dots are called *Feigenbaum cascades*.

Exercises

1. The Van der Pol equation (16.2) with $\varepsilon = 1$ possesses a limit cycle of period $T = 6.6632868593231301896996820305$ passing through $y_2 = 0$, $y_1 = A$ where $A = 2.00861986087484313650940188$. Replace (16.2) by

$$\begin{aligned}y_1' &= y_2(A - y_1) \\ y_2' &= ((1 - y_1^2)y_2 - y_1)(A - y_1)\end{aligned}$$

so that the limit cycle receives a stationary point. Study the behaviour of a solution starting in the interior, e.g. at $y_{10} = 1$, $y_{20} = 0$.

2. (Frommer 1934). Consider the system

$$y_1' = -y_2 + 2y_1y_2 - y_2^2, \quad y_2' = y_1 + (1 + \varepsilon)y_1^2 + 2y_1y_2 - y_2^2. \quad (16.24)$$

Show, either by a stability analysis similar to Exercise 5 of Section I.13 or by numerical computations, that for $\varepsilon > 0$ (16.24) possesses a limit cycle of asymptotic radius $r = \sqrt{6\varepsilon/7}$. (See also Wanner (1983), p. 15 and I.13, Exercise 5).

3. Solve Hilbert's 16th Problem: what is the highest possible number of limit cycles that a quadratic system

$$\begin{aligned}y_1' &= \alpha_0 + \alpha_1y_1 + \alpha_2y_2 + \alpha_3y_1^2 + \alpha_4y_1y_2 + \alpha_5y_2^2 \\ y_2' &= \beta_0 + \beta_1y_1 + \beta_2y_2 + \beta_3y_1^2 + \beta_4y_1y_2 + \beta_5y_2^2\end{aligned}$$

can have? The mathematical community is waiting for *you*: nobody has been able to solve this problem for more than 80 years. At the moment, the highest known number is 4, as for example in the system

$$\begin{aligned}y_1' &= \lambda y_1 - y_2 - 10y_1^2 + (5 + \delta)y_1y_2 + y_2^2 \\ y_2' &= y_1 + y_1^2 + (-25 + 8\varepsilon - 9\delta)y_1y_2, \\ \delta &= -10^{-13}, \quad \varepsilon = -10^{-52}, \quad \lambda = -10^{-200}\end{aligned}$$

(see Shi Songling 1980, Wanner 1983, Perko 1984).

4. Find a change of coordinates such that the equation

$$my'' + (-A + B(y')^2)y' + ky = 0$$

becomes the Van der Pol equation (16.2) (see Kryloff & Bogoliuboff (1947), p. 5).

5. Treat the pendulum equation

$$y'' + \sin y = y'' + y - \frac{y^3}{6} + \frac{y^5}{120} \pm \dots = 0, \quad y(0) = \varepsilon, \quad y'(0) = 0,$$

by the method of asymptotic expansions (16.6) and (16.7) and study the period as a function of ε .

Result. The period is $2\pi(1 + \varepsilon^2/16 + \dots)$.

6. Compute the limit cycle (Hopf bifurcation) for

$$y'' + y = \varepsilon^2 y' - (y')^3$$

for ε small by the method of Poincaré (16.6), (16.7) with $z'(0) = 0$.

7. Treat in a similar way as in Exercise 6 the Brusselator (16.12) with $A = 1$ and $B = 2 + \varepsilon^2$.

Hint. With the new variable $y = y_1 + y_2 - 3$ the differential equation (16.12) becomes equivalent to $y' = 1 - y_1$ and

$$y'' + y = -\varepsilon^2(y' - 1) - (y')^2(y + y') + 2yy'.$$

Result. $z(t) = \varepsilon(2/\sqrt{3}) \cos t + \dots$, $t = x(1 - \varepsilon^2/18 + \dots)$, so that the period is asymptotically $2\pi(1 + \varepsilon^2/18 + \dots)$.

8. (Liénard 1928). Prove that the limit cycle of the Van der Pol equation (16.1) is *unique* for every $\varepsilon > 0$.

Hint. The identity

$$y'' + \varepsilon(y^2 - 1)y' = \frac{d}{dx} \left(y' + \varepsilon \left(\frac{y^3}{3} - y \right) \right)$$

suggests the use of the coordinate system $y_1(x) = y(x)$, $y_2(x) = y' + \varepsilon(y^3/3 - y)$. Write the resulting first order system, study the signs of y'_1 , y'_2 and the increase of the “energy” function $V(x) = (y_1^2 + y_2^2)/2$.

Also generalize the result to equations of the form $y'' + f(y)y' + g(y) = 0$. For more details see e.g. Simmons (1972), p. 349.

9. (Rayleigh 1883). Compute the periodic solution of

$$y'' + \kappa y' + \lambda(y')^3 + n^2 y = 0$$

for κ and λ small.

Result. $y = A \sin(nx) + (\lambda n A^3/32) \cos(3nx) + \dots$ where A is given by $\kappa + (3/4)\lambda n^2 A^2 = 0$.

10. (Bendixson 1901). If in a certain region Ω of the plane the expression

$$\frac{\partial f_1}{\partial y_1} + \frac{\partial f_2}{\partial y_2}$$

is always negative or always positive, then the system (16.4) cannot have closed solutions in Ω .

Hint. Apply Green's formula

$$\int \int \left(\frac{\partial f_1}{\partial y_1} + \frac{\partial f_2}{\partial y_2} \right) dy_1 dy_2 = \int \left(f_1 dy_2 - f_2 dy_1 \right).$$

Chapter II. Runge-Kutta and Extrapolation Methods

Numerical methods for ordinary differential equations fall naturally into two classes: those which use *one* starting value at each step (“one-step methods”) and those which are based on *several* values of the solution (“multistep methods” or “multi-value methods”). The present chapter is devoted to the study of one-step methods, while multistep methods are the subject of Chapter III. Both chapters can, to a large extent, be read independently of each other.

We start with the theory of Runge-Kutta methods: the derivation of order conditions with the help of labelled trees, error estimates, convergence proofs, implementation, methods of higher order, dense output. Section II.7 introduces implicit Runge-Kutta methods. More attention will be drawn to these methods in Volume II on stiff differential equations. Two sections then discuss the elegant idea of *extrapolation* (Richardson, Romberg, etc) and its use in obtaining high order codes. The methods presented are then tested and compared on a series of problems. The potential of parallelism is discussed in a separate section. We then turn our attention to an algebraic theory of the composition of methods. This will be the basis for the study of order properties for many general classes of methods in the following chapter. The chapter ends with special methods for second order differential equations $y'' = f(x, y)$, for Hamiltonian systems (symplectic methods) and for problems with delay.

We illustrate the methods of this chapter with an example from Astronomy, the restricted three body problem. One considers two bodies of masses $1 - \mu$ and μ in circular rotation in a plane and a third body of negligible mass moving around in the same plane. The equations are (see e.g., the classical textbook Szebehely 1967)

$$\begin{aligned} y_1'' &= y_1 + 2y_2' - \mu' \frac{y_1 + \mu}{D_1} - \mu \frac{y_1 - \mu'}{D_2}, \\ y_2'' &= y_2 - 2y_1' - \mu' \frac{y_2}{D_1} - \mu \frac{y_2}{D_2}, \\ D_1 &= ((y_1 + \mu)^2 + y_2^2)^{3/2}, \quad D_2 = ((y_1 - \mu')^2 + y_2^2)^{3/2}, \\ \mu &= 0.012277471, \quad \mu' = 1 - \mu. \end{aligned} \tag{0.1}$$

There exist initial values, for example

$$\begin{aligned} y_1(0) &= 0.994, & y_1'(0) &= 0, & y_2(0) &= 0, \\ y_2'(0) &= -2.00158510637908252240537862224, \\ x_{\text{end}} &= 17.0652165601579625588917206249, \end{aligned} \quad (0.2)$$

such that the solution is periodic with period x_{end} . Such periodic solutions have fascinated astronomers and mathematicians for many decades (Poincaré; extensive numerical calculations are due to Sir George Darwin (1898)) and are now often called “Arenstorf orbits” (see Arenstorf (1963) who did numerical computations “on high speed electronic computers”). The problem is C^∞ with the exception of the two singular points $y_1 = -\mu$ and $y_1 = 1 - \mu$, $y_2 = 0$, therefore the Euler polygons of Section I.7 are known to converge to the solution. But are they really numerically useful here? We have chosen 24000 steps of step length $h = x_{\text{end}}/24000$ and plotted the result in Figure 0.1. The result is not very striking.

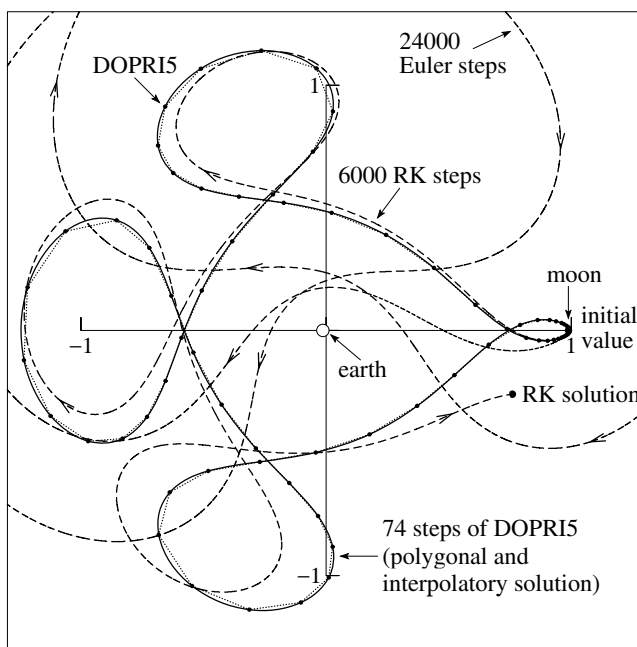


Fig. 0.1. An Arenstorf orbit computed by equidistant Euler, equidistant Runge-Kutta and variable step size Dormand & Prince

The performance of the Runge-Kutta method (left tableau of Table 1.2) is already much better and converges faster to the solution. We have used 6000 steps of step size $x_{\text{end}}/6000$, so that the numerical work becomes equivalent. Clearly, most accuracy is lost in those parts of the orbit which are close to a singularity. Therefore, codes with automatic step size selection, described in Section II.4, perform

much better and the code DOPRI5 (Table 5.2) computes the orbit with a precision of 10^{-3} in 98 steps (74 accepted and 24 rejected). The step size becomes very large in some regions and the graphical representation as polygons connecting the solution points becomes unsatisfactory. The solid line is the interpolatory solution (Section II.6), which is also precise for all intermediate values and useful for many other questions such as delay differential equations, event location or discontinuities in the differential equation.

For still higher precision one needs methods of higher order. For example, the code DOP853 (Section II.5) computes the orbit faster than DOPRI5 for more stringent tolerances, say smaller than about 10^{-6} . The highest possible order is obtained by extrapolation methods (Section II.9) and the code ODEX (with $K_{\max} = 15$) obtains the orbit with a precision of 10^{-30} with about 25000 function evaluations, precisely the same amount of work as for the above Euler solution.

II.1 The First Runge-Kutta Methods

Die numerische Berechnung irgend einer Lösung einer gegebenen Differentialgleichung, deren analytische Lösung man nicht kennt, hat, wie es scheint, die Aufmerksamkeit der Mathematiker bisher wenig in Anspruch genommen . . . (C. Runge 1895)

The Euler method for solving the initial value problem

$$y' = f(x, y), \quad y(x_0) = y_0 \quad (1.1)$$

was described by Euler (1768) in his “*Institutiones Calculi Integralis*” (Sectio Secunda, Caput VII). The method is easy to understand and to implement. We have studied its convergence extensively in Section I.7 and have seen that the global error behaves like Ch , where C is a constant depending on the problem and h is the maximal step size. If one wants a precision of, say, 6 decimals, one would thus need about a million steps, which is not very satisfactory. On the other hand, one knows since the time of Newton that much more accurate methods can be found, if f in (1.1) is independent of y , i.e., if we have a quadrature problem

$$y' = f(x), \quad y(x_0) = y_0 \quad (1.1')$$

with solution

$$y(X) = y_0 + \int_{x_0}^X f(x) dx. \quad (1.2)$$

As an example consider the midpoint rule (or first Gauss formula)

$$\begin{aligned} y(x_0 + h_0) &\approx y_1 = y_0 + h_0 f\left(x_0 + \frac{h_0}{2}\right) \\ y(x_1 + h_1) &\approx y_2 = y_1 + h_1 f\left(x_1 + \frac{h_1}{2}\right) \end{aligned} \quad (1.3')$$

...

$$y(X) \approx Y = y_{n-1} + h_{n-1} f\left(x_{n-1} + \frac{h_{n-1}}{2}\right),$$

where $h_i = x_{i+1} - x_i$ and $x_0, x_1, \dots, x_{n-1}, x_n = X$ is a subdivision of the integration interval. Its global error $y(X) - Y$ is known to be bounded by Ch^2 . Thus for a desired precision of 6 decimals, a thousand steps will usually do, i.e., the method here is a thousand times faster. Therefore Runge (1895) asked whether it would also be possible to extend method (1.3') to problem (1.1). The first step with $h = h_0$ would read

$$y(x_0 + h) \approx y_0 + hf\left(x_0 + \frac{h}{2}, y\left(x_0 + \frac{h}{2}\right)\right), \quad (1.3)$$

but which value should we take for $y(x_0 + h/2)$? In the absence of something better, it is natural to use one small Euler step with step size $h/2$ and obtain from (1.3) ¹

$$\begin{aligned} k_1 &= f(x_0, y_0) \\ k_2 &= f\left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2}k_1\right) \\ y_1 &= y_0 + hk_2. \end{aligned} \tag{1.4}$$

One might of course be surprised that we propose an Euler step for the computation of k_2 , just half a page after preaching its inefficiency. The crucial point is, however, that k_2 is multiplied by h in the third expression and therefore its error becomes less important. To be more precise, we compute the Taylor expansion of y_1 in (1.4) as a function of h ,

$$\begin{aligned} y_1 &= y_0 + hf\left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2}f_0\right) \\ &= y_0 + hf(x_0, y_0) + \frac{h^2}{2}(f_x + f_y f)(x_0, y_0) \\ &\quad + \frac{h^3}{8}(f_{xx} + 2f_{xy}f + f_{yy}f^2)(x_0, y_0) + \dots \end{aligned} \tag{1.5}$$

This can be compared with the Taylor series of the exact solution, which is obtained from (1.1) by repeated differentiation and replacing y' by f every time it appears (Euler (1768), Problema 86, §656, see also (8.12) of Chap. I)

$$\begin{aligned} y(x_0 + h) &= y_0 + hf(x_0, y_0) + \frac{h^2}{2}(f_x + f_y f)(x_0, y_0) \\ &\quad + \frac{h^3}{6}(f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_y f_x + f_y^2 f)(x_0, y_0) + \dots \end{aligned} \tag{1.6}$$

Subtracting these two equations, we obtain for the error of the first step

$$y(x_0 + h) - y_1 = \frac{h^3}{24}(f_{xx} + 2f_{xy}f + f_{yy}f^2 + 4(f_y f_x + f_y^2 f))(x_0, y_0) + \dots \tag{1.7}$$

When all second partial derivatives of f are bounded, we thus obtain

$$\|y(x_0 + h) - y_1\| \leq Kh^3.$$

In order to obtain an approximation of the solution of (1.1) at the endpoint X , we apply formula (1.4) successively to the intervals (x_0, x_1) , (x_1, x_2) , \dots , (x_{n-1}, X) , very similarly to the application of Euler's method in Section I.7. Again similarly to the convergence proof of Section I.7, it will be shown in Section II.3 that, as in the case (1.1'), the error of the numerical solution is bounded by Ch^2 (h the maximal step size). Method (1.4) is thus an improvement on the Euler method. For high precision computations we need to find still better methods; this will be the main task of what follows.

¹ The analogous extension of the *trapezoidal rule* has been given in an early publication by Coriolis in 1837; see Chapter II.4.2 of the thesis of D. Tournès, Paris VII, 1996.

General Formulation of Runge-Kutta Methods

Runge (1895) and Heun (1900) constructed methods by including additional Euler steps in (1.4). It was Kutta (1901) who then formulated the general scheme of what is now called a Runge-Kutta method:

Definition 1.1. Let s be an integer (the “number of stages”) and $a_{21}, a_{31}, a_{32}, \dots, a_{s1}, a_{s2}, \dots, a_{s,s-1}, b_1, \dots, b_s, c_2, \dots, c_s$ be real coefficients. Then the method

$$\begin{aligned}
 k_1 &= f(x_0, y_0) \\
 k_2 &= f(x_0 + c_2 h, y_0 + h a_{21} k_1) \\
 k_3 &= f(x_0 + c_3 h, y_0 + h(a_{31} k_1 + a_{32} k_2)) \\
 &\dots \\
 k_s &= f(x_0 + c_s h, y_0 + h(a_{s1} k_1 + \dots + a_{s,s-1} k_{s-1})) \\
 y_1 &= y_0 + h(b_1 k_1 + \dots + b_s k_s)
 \end{aligned} \tag{1.8}$$

is called an s -stage explicit Runge-Kutta method (ERK) for (1.1).

Usually, the c_i satisfy the conditions

$$c_2 = a_{21}, \quad c_3 = a_{31} + a_{32}, \quad \dots \quad c_s = a_{s1} + \dots + a_{s,s-1}, \tag{1.9}$$

or briefly,

$$c_i = \sum_{j=1}^{i-1} a_{ij}. \tag{1.9'}$$

These conditions, already assumed by Kutta, express that all points where f is evaluated are first order approximations to the solution. They greatly simplify the derivation of order conditions for high order methods. For low orders, however, these assumptions are not necessary (see Exercise 6).

Definition 1.2. A Runge-Kutta method (1.8) has *order* p if for sufficiently smooth problems (1.1),

$$\|y(x_0 + h) - y_1\| \leq K h^{p+1}, \tag{1.10}$$

i.e., if the Taylor series for the exact solution $y(x_0 + h)$ and for y_1 coincide up to (and including) the term h^p .

With the paper of Butcher (1964b) it became customary to symbolize method (1.8) by the tableau (1.8').

$$\begin{array}{c|ccc}
 0 & & & \\
 c_2 & a_{21} & & \\
 c_3 & a_{31} & a_{32} & \\
 \vdots & \vdots & \vdots & \ddots \\
 c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} \\
 \hline
 & b_1 & b_2 & \cdots & b_{s-1} & b_s
 \end{array} \quad (1.8')$$

Examples. The above method of Runge as well as methods of Runge and Heun of order 3 are given in Table 1.1.

Table 1.1. Low order Runge-Kutta methods

$ \begin{array}{c c} 0 & \\ 1/2 & 1/2 \\ \hline & 0 \quad 1 \\ \text{Runge, order 2} \end{array} $		$ \begin{array}{c ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 1 & 0 & 1 & \\ 1 & 0 & 0 & 1 \\ \hline & 1/6 & 2/3 & 0 & 1/6 \\ \text{Runge, order 3} \end{array} $				$ \begin{array}{c ccc} 0 & & & \\ 1/3 & 1/3 & & \\ 2/3 & 0 & 2/3 & \\ \hline & 1/4 & 0 & 3/4 \\ \text{Heun, order 3} \end{array} $			
--	--	--	--	--	--	---	--	--	--

Discussion of Methods of Order 4

Von den neueren Verfahren halte ich das folgende von Herrn Kutta angegebene für das beste. (C. Runge 1905)

Our task is now to determine the coefficients of 4-stage Runge-Kutta methods (1.8) in order that they be of order 4. We have seen above what we must do: compute the derivatives of $y_1 = y_1(h)$ for $h = 0$ and compare them with those of the true solution for orders 1, 2, 3, and 4. In theory, with the known rules of differential calculus, this is a completely trivial task and, by the use of (1.9), results in the following conditions:

$$\sum_i b_i = b_1 + b_2 + b_3 + b_4 = 1 \quad (1.11a)$$

$$\sum_i b_i c_i = b_2 c_2 + b_3 c_3 + b_4 c_4 = 1/2 \quad (1.11b)$$

$$\sum_i b_i c_i^2 = b_2 c_2^2 + b_3 c_3^2 + b_4 c_4^2 = 1/3 \quad (1.11c)$$

$$\sum_{i,j} b_i a_{ij} c_j = b_3 a_{32} c_2 + b_4 (a_{42} c_2 + a_{43} c_3) = 1/6 \quad (1.11d)$$

$$\sum_i b_i c_i^3 = b_2 c_2^3 + b_3 c_3^3 + b_4 c_4^3 = 1/4 \quad (1.11e)$$

$$\sum_{i,j} b_i c_i a_{ij} c_j = b_3 c_3 a_{32} c_2 + b_4 c_4 (a_{42} c_2 + a_{43} c_3) = 1/8 \quad (1.11f)$$

$$\sum_{i,j} b_i a_{ij} c_j^2 = b_3 a_{32} c_2^2 + b_4 (a_{42} c_2^2 + a_{43} c_3^2) = 1/12 \quad (1.11g)$$

$$\sum_{i,j,k} b_i a_{ij} a_{jk} c_k = b_4 a_{43} a_{32} c_2 = 1/24. \quad (1.11h)$$

These computations, which are not reproduced in Kutta's paper (they are, however, in Heun 1900), are very tedious. And they grow enormously with higher orders. We shall see in Section II.2 that by using an appropriate notation, they can become very elegant.

Kutta gave the general solution of (1.11) without comment. A clear derivation of the solutions is given in Runge & König (1924), p. 291. We shall follow here the ideas of J.C. Butcher, which make clear the role of the so-called *simplifying assumptions*, and will also apply to higher order cases.

Lemma 1.3. *If*

$$\sum_{i=j+1}^s b_i a_{ij} = b_j (1 - c_j), \quad j = 1, \dots, s, \quad (1.12)$$

then the equations (d), (g), and (h) in (1.11) follow from the others.

Proof. We demonstrate this for (g):

$$\sum_{i,j} b_i a_{ij} c_j^2 = \sum_j b_j c_j^2 - \sum_j b_j c_j^3 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

by (c) and (e). Equations (d) and (h) are derived similarly. \square

We shall now show that (1.12) is also *necessary* in our case:

Lemma 1.4. *For $s = 4$, the equations (1.11) and (1.9) imply (1.12).*

The proof of this lemma will be based on the following:

Lemma 1.5. *Let U and V be 3×3 matrices such that*

$$UV = \begin{pmatrix} a & b & 0 \\ c & d & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} \neq 0. \quad (1.13)$$

Then either $Ve_3 = 0$ or $U^T e_3 = 0$ where $e_3 = (0, 0, 1)^T$.

Proof of Lemma 1.5. If $\det U \neq 0$, then $UVe_3 = 0$ implies $Ve_3 = 0$. If $\det U = 0$, there exists $x = (x_1, x_2, x_3)^T \neq 0$ such that $U^T x = 0$, and therefore $V^T U^T x = 0$. But (1.13) implies that x must be a multiple of e_3 . \square

Proof of Lemma 1.4. Define

$$d_j = \sum_i b_i a_{ij} - b_j(1 - c_j) \quad \text{for} \quad j = 1, \dots, 4,$$

so that we have to prove $d_j = 0$. We now introduce the matrices

$$U = \begin{pmatrix} b_2 & b_3 & b_4 \\ b_2 c_2 & b_3 c_3 & b_4 c_4 \\ d_2 & d_3 & d_4 \end{pmatrix}, \quad V = \begin{pmatrix} c_2 & c_2^2 & \sum_j a_{2j} c_j - c_2^2/2 \\ c_3 & c_3^2 & \sum_j a_{3j} c_j - c_3^2/2 \\ c_4 & c_4^2 & \sum_j a_{4j} c_j - c_4^2/2 \end{pmatrix}. \quad (1.14)$$

Multiplication of these two matrices, using the conditions of (1.11), gives

$$UV = \begin{pmatrix} 1/2 & 1/3 & 0 \\ 1/3 & 1/4 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{with} \quad \det \begin{pmatrix} 1/2 & 1/3 \\ 1/3 & 1/4 \end{pmatrix} \neq 0.$$

Now the last column of V cannot be zero, since $c_1 = 0$ implies

$$\sum_j a_{2j} c_j - c_2^2/2 = -c_2^2/2 \neq 0$$

by condition (h). Thus $d_2 = d_3 = d_4 = 0$ follows from Lemma 1.5. The last identity $d_1 = 0$ follows from $d_1 + d_2 + d_3 + d_4 = 0$, which is a consequence of (1.11a,b) and (1.9). \square

From Lemmas 1.3 and 1.4 we obtain

Theorem 1.6. *Under the assumption (1.9) the equations (1.11) are equivalent to*

$$b_1 + b_2 + b_3 + b_4 = 1 \quad (1.15a)$$

$$b_2 c_2 + b_3 c_3 + b_4 c_4 = 1/2 \quad (1.15b)$$

$$b_2 c_2^2 + b_3 c_3^2 + b_4 c_4^2 = 1/3 \quad (1.15c)$$

$$b_2 c_2^3 + b_3 c_3^3 + b_4 c_4^3 = 1/4 \quad (1.15e)$$

$$b_3 c_3 a_{32} c_2 + b_4 c_4 (a_{42} c_2 + a_{43} c_3) = 1/8 \quad (1.15f)$$

$$b_3 a_{32} + b_4 a_{42} = b_2(1 - c_2) \quad (1.15i)$$

$$b_4 a_{43} = b_3(1 - c_3) \quad (1.15j)$$

$$0 = b_4(1 - c_4). \quad (1.15k)$$

\square

It follows from (1.15j) and (1.11h) that

$$b_3 b_4 c_2 (1 - c_3) \neq 0. \quad (1.16)$$

In particular this implies $c_4 = 1$ by (1.15k).

Solution of equations (1.15). Equations (a)-(e) and (k) just state that b_i and c_i are the coefficients of a fourth order quadrature formula with $c_1 = 0$ and $c_4 = 1$. We distinguish four cases for this:

$$1) \quad c_2 = u, \quad c_3 = v \quad \text{and} \quad 0, u, v, 1 \text{ are all distinct}; \quad (1.17)$$

then (a)-(e) form a regular linear system for b_1, b_2, b_3, b_4 . This system has the solution

$$\begin{aligned} b_1 &= \frac{1 - 2(u + v) + 6uv}{12uv}, & b_2 &= \frac{2v - 1}{12u(1 - u)(v - u)}, \\ b_3 &= \frac{1 - 2u}{12v(1 - v)(v - u)}, & b_4 &= \frac{3 - 4(u + v) + 6uv}{12(1 - u)(1 - v)}. \end{aligned}$$

Due to (1.16) we have to assume that u, v are such that $b_3 \neq 0$ and $b_4 \neq 0$. The three other cases with double nodes are built upon the Simpson rule:

$$2) \quad c_3 = 0, \quad c_2 = 1/2, \quad b_3 = w \neq 0, \quad b_1 = 1/6 - w, \quad b_2 = 4/6, \quad b_4 = 1/6;$$

$$3) \quad c_2 = c_3 = 1/2, \quad b_1 = 1/6, \quad b_3 = w \neq 0, \quad b_2 = 4/6 - w, \quad b_4 = 1/6;$$

$$4) \quad c_2 = 1, \quad c_3 = 1/2, \quad b_4 = w \neq 0, \quad b_2 = 1/6 - w, \quad b_1 = 1/6, \quad b_3 = 4/6.$$

Once b_i and c_i are chosen, we obtain a_{43} from (j), and then (f) and (i) form a linear system of two equations for a_{32} and a_{42} . The determinant of this system is

$$\det \begin{pmatrix} b_3 & b_4 \\ b_3 c_3 c_2 & b_4 c_4 c_2 \end{pmatrix} = b_3 b_4 c_2 (c_4 - c_3)$$

which is $\neq 0$ by (1.16). Finally we obtain a_{21} , a_{31} , and a_{41} from (1.9).

Two particular choices of Kutta (1901) have become especially popular: case (3) with $w = 2/6$ and case (1) with $u = 1/3$, $v = 2/3$. They are given in Table 1.2. Both methods generalize classical quadrature rules in keeping the same order. The first is more popular, the second is more precise ("Wir werden diese Näherung als im allgemeinen beste betrachten . . .", Kutta).

Table 1.2. Kutta's methods

0					0				
1/2	1/2				1/3	1/3			
1/2	0	1/2			2/3	-1/3	1		
1	0	0	1		1	1	-1	1	

“Optimal” Formulas

Much research has been undertaken, in order to choose the “best” possibilities from the variety of possible 4th order RK-formulas.

The first attempt in this direction was the very popular method of Gill (1951), with the aim of reducing the need for computer storage (“registers”) as much as possible. The first computers in the fifties largely used this method which is therefore of historical interest. Gill observed that most computer storage is needed for the computation of k_3 , where “registers are required to store in some form”

$$y_0 + a_{31}hk_1 + a_{32}hk_2, \quad y_0 + a_{41}hk_1 + a_{42}hk_2, \quad y_0 + b_1hk_1 + b_2hk_2, \quad hk_3.$$

“Clearly, three registers will suffice for the third stage if the quantities to be stored are linearly dependent, i.e., if”

$$\det \begin{pmatrix} 1 & a_{31} & a_{32} \\ 1 & a_{41} & a_{42} \\ 1 & b_1 & b_2 \end{pmatrix} = 0.$$

Gill observed that this condition is satisfied for the methods of type (3) if $w = (1 + \sqrt{0.5})/3$. The resulting method can then be reformulated as follows (“As each quantity is calculated it is stored in the register formerly holding the corresponding quantity of the previous stage, which is no longer required”):

$$\begin{aligned} y &:= \text{initial value}, & k &:= hf(y), & y &:= y + 0.5k, & q &:= k, \\ k &:= hf(y), & y &:= y + (1 - \sqrt{0.5})(k - q), \\ q &:= (2 - \sqrt{2})k + (-2 + 3\sqrt{0.5})q, \\ k &:= hf(y), & y &:= y + (1 + \sqrt{0.5})(k - q), \\ q &:= (2 + \sqrt{2})k + (-2 - 3\sqrt{0.5})q, \\ k &:= hf(y), & y &:= y + \frac{k}{6} - \frac{q}{3}, & (\rightarrow \text{ compute next step}) . \end{aligned} \tag{1.18}$$

Today, in large high-speed computers, this method is no longer used, but could still be of interest for very high dimensional equations.

Other attempts have been made to choose u and v in (1.17), case (1), such that the *error terms* (terms in h^5 , see Section II.3) become as small as possible. We shall discuss this question in Section II.3.

Numerical Example

Zu grosses Gewicht darf man natürlich solchen Beispielen nicht
beilegen . . .
(W. Kutta 1901)

We compare five different choices of 4th order methods on the Van der Pol equation (I.16.2) with $\varepsilon = 1$. As initial values we take $y_1(0) = A$, $y_2(0) = 0$ on the limit cycle and we integrate over one period T (the values of A and T are given in Exercise I.16.1). For a comparison of these methods with lower order ones we have also included the explicit Euler method, Runge's method of order 2 and Heun's method of order 3 (see Table 1.1).

We have applied the methods with several fixed step sizes. The errors of both components and the number of function evaluations (fe) are displayed in logarithmic scales in Fig. 1.1. Whenever the error behaves like $C \cdot h^p = C_1 \cdot (fe)^{-p}$, the curves appear as straight lines with slope $1/p$. We have chosen the scales such that the theoretical slope of the 4th order methods appears to be 45° .

These tests clearly show up the importance of higher order methods. Among the various 4th order methods there is usually no big difference. It is interesting to note that in our example the method with the smallest error in y_1 has the biggest error in y_2 and vice versa.

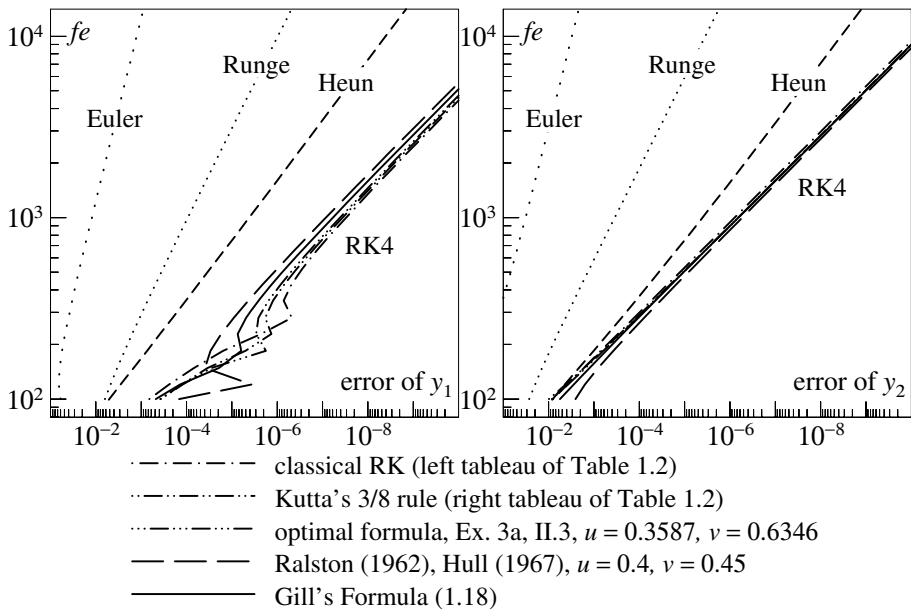


Fig. 1.1. Global errors versus number of function evaluations

Exercises

1. Show that every s -stage explicit RK method of order s , when applied to the problem $y' = \lambda y$ (λ a complex constant), gives

$$y_1 = \left(\sum_{j=0}^s \frac{z^j}{j!} \right) y_0, \quad z = h\lambda.$$

Hint. Show first that y_1/y_0 must be a polynomial in z of degree s and then determine its coefficients by comparing the derivatives of y_1 , with respect to h , to those of the true solution.

2. (Runge 1895, p. 175; see also the introduction to Adams methods in Chap. III.1). The theoretical form of drops of fluids is determined by the differential equation of Laplace (1805)

$$-z = \alpha^2 \frac{(K_1 + K_2)}{2} \quad (1.21)$$

where α is a constant, $(K_1 + K_2)/2$ the mean curvature, and z the height (see Fig. 1.2). If we insert $1/K_1 = r/\sin \varphi$ and $K_2 = d\varphi/ds$, the curvature of the meridian curve, we obtain

$$-2z = \alpha^2 \left(\frac{\sin \varphi}{r} + \frac{d\varphi}{ds} \right), \quad (1.22)$$

where we put $\alpha = 1$. Add

$$\frac{dr}{ds} = \cos \varphi, \quad \frac{dz}{ds} = -\sin \varphi, \quad (1.22')$$

to obtain a system of three differential equations for $\varphi(s)$, $r(s)$, $z(s)$, s being the arc length. Compute and plot different solution curves by the method of Runge (1.4) with initial values $\varphi(0) = 0$, $r(0) = 0$ and $z(0) = z_0$ ($z_0 < 0$ for lying drops; compute also hanging drops with appropriate sign changes in (1.22)). Use different step sizes and compare the results.

Hint. Be careful at the singularity in the beginning: from (1.22) and (1.22') we have for small s that $r = s$, $\varphi = \zeta s$ with $\zeta = -z_0$, hence $(\sin \varphi)/r \rightarrow -z_0$. A more precise analysis gives for small s the expansions ($\zeta = -z_0$)

$$\begin{aligned} \varphi &= \zeta s + \frac{\zeta}{4} s^3 + \left(\frac{\zeta}{48} - \frac{\zeta^3}{120} \right) s^5 + \dots \\ r &= s - \frac{\zeta^2}{6} s^3 + \left(-\frac{\zeta^2}{20} + \frac{\zeta^4}{120} \right) s^5 + \dots \\ z &= -\zeta - \frac{\zeta}{2} s^2 + \left(-\frac{\zeta}{16} + \frac{\zeta^3}{24} \right) s^4 + \left(-\frac{\zeta}{288} + \frac{\zeta^3}{45} - \frac{\zeta^5}{720} \right) s^6 + \dots \end{aligned}$$

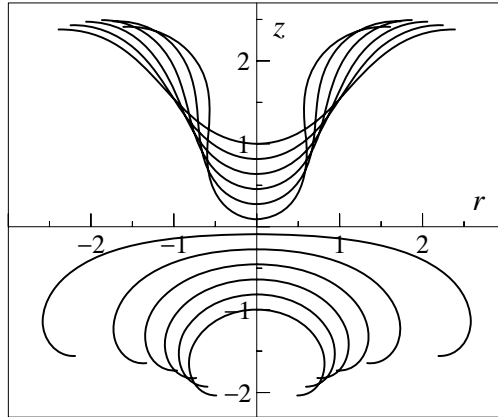


Fig. 1.2. Drops

3. Find the conditions for a 2-stage explicit RK-method to be of order two and determine all such methods (“... wozu eine weitere Erörterung nicht mehr nötig ist”, Kutta).
4. Find all methods of order three with three stages (i.e., solve (1.11;a-d) with $b_4 = 0$).

Result. $c_2 = u$, $c_3 = v$, $a_{32} = v(v - u)/(u(2 - 3u))$, $b_2 = (2 - 3v)/(6u(u - v))$, $b_3 = (2 - 3u)/(6v(v - u))$, $b_1 = 1 - b_2 - b_3$, $a_{31} = c_3 - a_{32}$, $a_{21} = c_2$ (Kutta 1901, p. 438).

5. Construct all methods of order 2 of the form

$$\begin{array}{c|cc} 0 & & \\ c_2 & c_2 & \\ c_3 & 0 & c_3 \\ \hline & 0 & 0 & 1 \end{array}$$

Such methods “have the property that the corresponding Runge-Kutta process requires relatively less storage in a computer” (Van der Houwen (1977), §2.7.2). Apply them to $y' = \lambda y$ and compare with Exercise 1.

6. Determine the conditions for order two of the RK methods with two stages which do *not* satisfy the conditions (1.9):

$$\begin{aligned} k_1 &= f(x_0 + c_1 h, y_0) \\ k_2 &= f(x_0 + c_2 h, y_0 + a_{21} h k_1) \\ y_1 &= y_0 + h(b_1 k_1 + b_2 k_2). \end{aligned}$$

Discuss the use of this extra freedom for c_1 and c_2 (Oliver 1975).

II.2 Order Conditions for Runge-Kutta Methods

... I heard a lecture by Merson ...
(J. Butcher's first contact with RK methods)

In this section we shall derive the general structure of the order conditions (Merson 1957, Butcher 1963). The proof has evolved very much in the meantime, mainly under the influence of Butcher's later work, many personal discussions with him, the proof of "Theorem 6" in Hairer & Wanner (1974), and our teaching experience. We shall see in Section II.11 that exactly the same ideas of proof lead to a general theorem of composition of methods (= *B*-series), which gives access to order conditions for a much larger class of methods.

A big advantage is obtained by transforming (1.1) to *autonomous* form by appending x to the dependent variables as

$$\begin{pmatrix} x \\ y \end{pmatrix}' = \begin{pmatrix} 1 \\ f(x, y) \end{pmatrix}. \quad (2.1)$$

The main difficulty in the derivation of the order conditions is to understand the correspondence of the formulas to certain rooted labelled trees; this comes out most naturally if we use well-chosen indices and tensor notation (as in Gill (1951), Henrici (1962), p. 118, Gear (1971), p. 32). As is usual in tensor notation, we denote (in this section) the components of vectors by *superscript* indices which, in order to avoid confusion, we choose as *capitals*. Then (2.1) can be written as

$$(y^J)' = f^J(y^1, \dots, y^n), \quad J = 1, \dots, n. \quad (2.2)$$

We next rewrite the method (1.8) for the autonomous differential equation (2.2). In order to get a better symmetry in all formulas of (1.8), we replace k_i by the argument g_i such that $k_i = f(g_i)$. Then (1.8) becomes

$$\begin{aligned} g_i^J &= y_0^J + \sum_{j=1}^{i-1} a_{ij} h f^J(g_j^1, \dots, g_j^n), \quad i = 1, \dots, s \\ y_1^J &= y_0^J + \sum_{j=1}^s b_j h f^J(g_j^1, \dots, g_j^n). \end{aligned} \quad (2.3)$$

If the system (2.2) originates from (2.1), then, for $J = 1$,

$$g_i^1 = y_0^1 + \sum_{j=1}^{i-1} a_{ij} h = x_0 + c_i h$$

by (1.9). We see that (1.9) becomes a natural condition. If it is satisfied, then for the derivation of order conditions only the autonomous equation (2.2) has to be considered.

As indicated in Section II.1 we have to compare the Taylor series of y_1^J with that of the exact solution. Therefore we compute the derivatives of y_1^J and g_i^J with respect to h at $h = 0$. Due to the similarity of the two formulas, it is sufficient to do this for g_i^J . On the right hand side of (2.3) there appear expressions of the form $h\varphi(h)$, so we make use of Leibniz' formula

$$(h\varphi(h))^{(q)}|_{h=0} = q \cdot (\varphi(h))^{(q-1)}|_{h=0}. \quad (2.4)$$

The reader is now asked to take a deep breath, take five sheets of reversed computer paper, remember the basic rules of differential calculus, and begin the following computations:

$q = 0$: from (2.3)

$$(g_i^J)^{(0)}|_{h=0} = y_0^J. \quad (2.5;0)$$

$q = 1$: from (2.3) and (2.4)

$$(g_i^J)^{(1)}|_{h=0} = \sum_j a_{ij} f^J|_{y=y_0}. \quad (2.5;1)$$

$q = 2$: because of (2.4) we shall need the first derivative of $f^J(g_j)$

$$(f^J(g_j))^{(1)} = \sum_K f_K^J(g_j) \cdot (g_j^K)^{(1)}, \quad (2.6;1)$$

where, as usual, f_K^J denotes $\partial f^J / \partial y^K$. Inserting formula (2.5;1) (with i, j, J replaced by j, k, K) into (2.6;1) we obtain with (2.4)

$$(g_i^J)^{(2)}|_{h=0} = 2 \sum_{j,k} a_{ij} a_{jk} \sum_K f_K^J f^K|_{y=y_0}. \quad (2.5;2)$$

$q = 3$: we differentiate (2.6;1) to obtain

$$(f^J(g_j))^{(2)} = \sum_{K,L} f_{KL}^J(g_j) \cdot (g_j^K)^{(1)} (g_j^L)^{(1)} + \sum_K f_K^J(g_j) (g_j^K)^{(2)}. \quad (2.6;2)$$

The derivatives $(g_j^K)^{(1)}$ and $(g_j^K)^{(2)}$ at $h = 0$ are already available in (2.5;1) and (2.5;2). So we have from (2.3) and (2.4)

$$\begin{aligned} (g_i^J)^{(3)}|_{h=0} &= 3 \sum_{j,k,l} a_{ij} a_{jk} a_{jl} \sum_{K,L} f_{KL}^J f^K f^L|_{y=y_0} \\ &\quad + 3 \cdot 2 \sum_{j,k,l} a_{ij} a_{jk} a_{kl} \sum_{K,L} f_K^J f_L^K f^L|_{y=y_0}. \end{aligned} \quad (2.5;3)$$

The same formula holds for $(y_1^J)^{(3)}|_{h=0}$ with a_{ij} replaced by b_j .

The Derivatives of the True Solution

The derivatives of the correct solution are obtained much more easily just by differentiating equation (2.2): first

$$(y^J)^{(1)} = f^J(y). \quad (2.7;1)$$

Differentiating (2.2) and inserting (2.2) again for the derivatives we get

$$(y^J)^{(2)} = \sum_K f_K^J(y) \cdot (y^K)^{(1)} = \sum_K f_K^J(y) f^K(y). \quad (2.7;2)$$

Differentiating (2.7;2) again we obtain

$$(y^J)^{(3)} = \sum_{K,L} f_{KL}^J(y) f^K(y) f^L(y) + \sum_{K,L} f_K^J(y) f_L^K(y) f^L(y). \quad (2.7;3)$$

Conditions for Order 3

For order 3, the derivatives (2.5;1-3), (with a_{ij} replaced by b_j) must be equal to the derivatives (2.7;1-3), and this for every differential equation. Thus, comparing the corresponding expressions, we obtain:

Theorem 2.1. *The RK method (2.3) (and thus (1.8)) is of order 3 iff*

$$\begin{aligned} \sum_j b_j &= 1, & 2 \sum_{j,k} b_j a_{jk} &= 1, \\ 3 \sum_{j,k,l} b_j a_{jk} a_{jl} &= 1, & 6 \sum_{j,k,l} b_j a_{jk} a_{kl} &= 1. \end{aligned} \quad (2.8)$$

□

Inserting $\sum_k a_{jk} = c_j$ from (1.9), we can simplify these expressions still further and obtain formulas (a)-(d) of (1.11).

Trees and Elementary Differentials

But without a more convenient notation, it would be difficult to find the corresponding expressions . . . This, however, can be at once effected by means of the analytical forms called trees . . .

(A. Cayley 1857)

The continuation of this process, although theoretically clear, soon leads to very complicated formulas. It is therefore advantageous to use a graphical representation: indeed, the indices j, k, l and J, K, L in the terms of (2.5;3) are linked

together as pairs of indices in a_{jk}, a_{jl}, \dots in exactly the same way as upper and lower indices in the expressions f_{KL}^J, f_K^J , namely

$$t_{31} = \begin{array}{c} l \quad k \\ \diagdown \quad \diagup \\ j \end{array} \quad \text{and} \quad t_{32} = \begin{array}{c} l \\ \diagdown \quad \diagup \\ j \quad k \end{array} \quad (2.9)$$

for the first and second term respectively. We call these objects *labelled trees*, because they are connected graphs (trees) whose vertices are labelled with summation indices. They can also be represented as *mappings*, e.g.,

$$l \mapsto j, \quad k \mapsto j \quad \text{and} \quad l \mapsto k, \quad k \mapsto j \quad (2.9')$$

for the above trees. This mapping indicates to which lower letter the corresponding vertices are attached.

Definition 2.2. Let A be an ordered chain of indices $A = \{j < k < l < m < \dots\}$ and denote by A_q the subset consisting of the first q indices. A (*rooted*) *labelled tree* of order q ($q \geq 1$) is a mapping (the son-father mapping)

$$t : A_q \setminus \{j\} \rightarrow A_q$$

such that $t(z) < z$ for all $z \in A_q \setminus \{j\}$. The set of all labelled trees of order q is denoted by LT_q . We call “ z ” the *son* of “ $t(z)$ ” and “ $t(z)$ ” the *father* of “ z ”. The vertex “ j ”, the forefather of the whole dynasty, is called the *root* of t . The order q of a labelled tree is equal to the number of its vertices and is usually denoted by $q = \varrho(t)$.

Definition 2.3. For a labelled tree $t \in LT_q$ we call

$$F^J(t)(y) = \sum_{K, L, \dots} f_{K, \dots}^J(y) f_{\dots}^K(y) f_{\dots}^L(y) \dots$$

the corresponding *elementary differential*. The summation is over $q - 1$ indices K, L, \dots (which correspond to $A_q \setminus \{j\}$) and the summand is a product of q f 's, where the upper index runs through all vertices of t and the lower indices are the corresponding sons. We denote by $F(t)(y)$ the vector $(F^1(t)(y), \dots, F^n(t)(y))$.

If the set A_q is written as

$$A_q = \{j_1 < j_2 < \dots < j_q\}, \quad (2.10)$$

then we can write the definition of $F(t)$ as follows:

$$F^{J_1}(t) = \sum_{J_2, \dots, J_q} \prod_{i=1}^q f_{t^{-1}(J_i)}^{J_i}, \quad (2.11)$$

since the sons of an index are its inverse images under the map t .

Examples of elementary differentials are

$$\sum_{K,L} f_{KL}^J f^K f^L \quad \text{and} \quad \sum_{K,L} f_K^J f_L^K f^L$$

for the labelled trees t_{31} and t_{32} above. These expressions appear in formulas (2.5;3) and (2.7;3).

The three labelled trees

$$\begin{array}{ccc} \begin{array}{c} l \\ \diagdown \quad \diagup \\ m \quad j \quad k \end{array} & \begin{array}{c} m \\ \diagdown \quad \diagup \\ l \quad j \quad k \end{array} & \begin{array}{c} m \\ \diagdown \quad \diagup \\ k \quad j \quad l \end{array} \end{array} \quad (2.12)$$

all look topologically alike, moreover the corresponding elementary differentials

$$\sum_{K,L,M} f_{KM}^J f^M f_L^K f^L, \quad \sum_{K,L,M} f_{KL}^J f^L f_M^K f^M, \quad \sum_{K,L,M} f_{LK}^J f^K f_M^L f^M \quad (2.12')$$

are the same, because they just differ by an exchange of the summation indices. Thus we give

Definition 2.4. Two labelled trees t and u are *equivalent*, if they have the same order, say q , and if there exists a permutation $\sigma : A_q \rightarrow A_q$, such that $\sigma(j) = j$ and $t\sigma = \sigma u$ on $A_q \setminus \{j\}$.

This clearly defines an equivalence relation.

Definition 2.5. An equivalence class of q th order labelled trees is called a (*rooted*) *tree of order q* . The set of all trees of order q is denoted by T_q . The *order* of a tree is defined as the order of a representative and is again denoted by $\varrho(t)$. Furthermore we denote by $\alpha(t)$ (for $t \in T_q$) the number of elements in the equivalence class t ; i.e., the number of possible different monotonic labellings of t .

Geometrically, a tree is distinguished from a labelled tree by omitting the labels. Often it is advantageous to include \emptyset , the empty tree, as the only tree of order 0. The only tree of order 1 is denoted by τ . The number of trees of orders $1, 2, \dots, 10$ are given in Table 2.1. Representatives of all trees of order ≤ 5 are shown in Table 2.2.

Table 2.1. Number of trees up to order 10

q	1	2	3	4	5	6	7	8	9	10
card(T_q)	1	1	2	4	9	20	48	115	286	719

Table 2.2. Trees and elementary differentials up to order 5

q	t	graph	$\gamma(t)$	$\alpha(t)$	$F^J(t)(y)$	$\Phi_j(t)$
0	\emptyset	\emptyset	1	1	y^J	
1	τ	$\bullet j$	1	1	f^J	1
2	t_{21}		2	1	$\sum_K f_K^J f^K$	$\sum_k a_{jk}$
3	t_{31}		3	1	$\sum_{K,L} f_{KL}^J f^K f^L$	$\sum_{k,l} a_{jk} a_{jl}$
	t_{32}		6	1	$\sum_{K,L} f_K^J f_L^K f^L$	$\sum_{k,l} a_{jk} a_{kl}$
4	t_{41}		4	1	$\sum_{K,L,M} f_{KLM}^J f^K f^L f^M$	$\sum_{k,l,m} a_{jk} a_{jl} a_{jm}$
	t_{42}		8	3	$\sum_{K,L,M} f_{KM}^J f_L^K f^L f^M$	$\sum_{k,l,m} a_{jk} a_{kl} a_{jm}$
	t_{43}		12	1	$\sum_{K,L,M} f_K^J f_{LM}^K f^L f^M$	$\sum_{k,l,m} a_{jk} a_{kl} a_{km}$
	t_{44}		24	1	$\sum_{K,L,M} f_K^J f_L^K f_M^L f^M$	$\sum_{k,l,m} a_{jk} a_{kl} a_{lm}$
5	t_{51}		5	1	$\sum f_{KLMP}^J f^K f^L f^M f^P$	$\sum a_{jk} a_{jl} a_{jm} a_{jp}$
	t_{52}		10	6	$\sum f_{KMP}^J f_L^K f^L f^M f^P$	$\sum a_{jk} a_{kl} a_{jm} a_{jp}$
	t_{53}		15	4	$\sum f_{KLP}^J f_M^K f^L f^M f^P$	$\sum a_{jk} a_{kl} a_{km} a_{jp}$
	t_{54}		30	4	$\sum f_{KLP}^J f_L^K f_M^L f^M f^P$	$\sum a_{jk} a_{kl} a_{lm} a_{jp}$
	t_{55}		20	3	$\sum f_{KM}^J f_L^K f^L f_P^M f^P$	$\sum a_{jk} a_{kl} a_{jm} a_{mp}$
	t_{56}		20	1	$\sum f_K^J f_{LMP}^K f^L f^M f^P$	$\sum a_{jk} a_{kl} a_{km} a_{kp}$
	t_{57}		40	3	$\sum f_K^J f_L^K f_{MP}^L f^M f^P$	$\sum a_{jk} a_{kl} a_{lm} a_{kp}$
	t_{58}		60	1	$\sum f_K^J f_L^K f_{MP}^L f^M f^P$	$\sum a_{jk} a_{kl} a_{lm} a_{lp}$
	t_{59}		120	1	$\sum f_K^J f_L^K f_M^L f_P^M f^P$	$\sum a_{jk} a_{kl} a_{lm} a_{mp}$

The Taylor Expansion of the True Solution

We can now state the general result for the q th derivative of the true solution:

Theorem 2.6. *The exact solution of (2.2) satisfies*

$$(y)^{(q)}(x_0) = \sum_{t \in LT_q} F(t)(y_0) = \sum_{t \in T_q} \alpha(t) F(t)(y_0). \quad (2.7;q)$$

Proof. The theorem is true for $q = 1, 2, 3$ (see (2.7;1-3) above). For the computation of, say, the 4th derivative, we have to differentiate (2.7;3). This consists of two terms (corresponding to the two trees of (2.9)), each of which contains three factors f, \dots (corresponding to the three nodes of these trees). The differentiation of these by Leibniz' rule and insertion of (2.2) for the derivatives is geometrically just the addition of a new branch with a new summation letter to *each* vertex (Fig. 2.1).

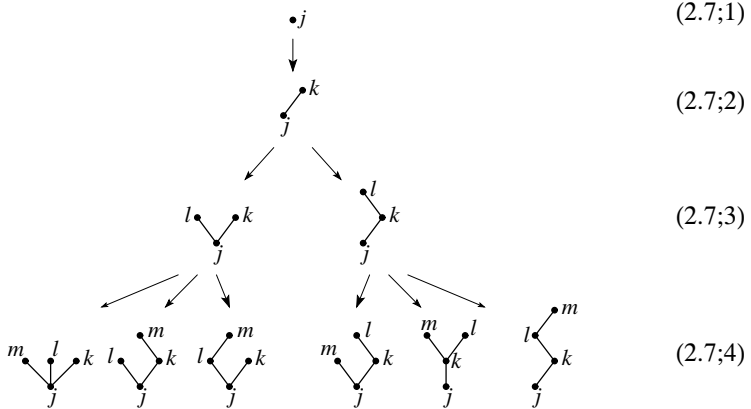


Fig. 2.1. Derivatives of exact solution

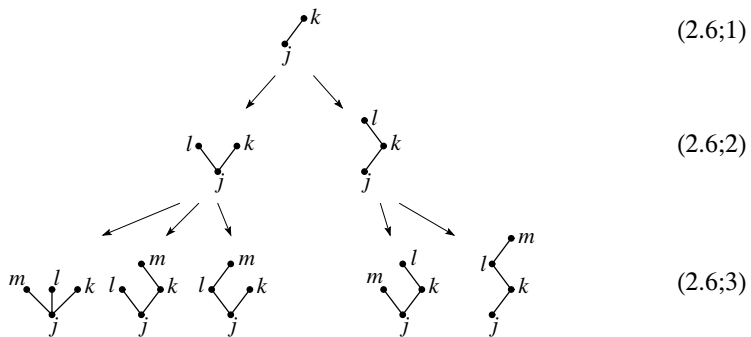
It is clear that by this process *all* labelled trees of order q appear for the q th derivative, each of them *exactly once*.

If we group together the terms with identical elementary differentials, we obtain the second expression of (2.7;q). □

Faà di Bruno's Formula

Our next goal will be the computation of the q th derivative of the numerical solution y_1 and of the g_j . For this, we have first to generalize the formulas (2.6;1) (the chain rule) and (2.6;2) for the q th derivative of the composition of two functions. We represent these two formulas graphically in Fig. 2.2.

Formula (2.6;2) consists of two terms; the first term contains three factors, the second contains only two. Here the node "l" is a "dummy" node, not really present in the formula, and just indicates that we have to take the second derivative. The derivation of (2.6;2) will thus lead to *five* terms which we write down for the convenience of the reader (but not for the convenience of the printer ...)


 Fig. 2.2. Derivatives of $f^J(g)$

$$\begin{aligned}
 (f^J(g))^{(3)} &= \sum_{K,L,M} f_{KLM}^J(g) \cdot (g^K)^{(1)}(g^L)^{(1)}(g^M)^{(1)} \\
 &+ \sum_{K,L} f_{KL}^J(g) \cdot (g^K)^{(2)}(g^L)^{(1)} + \sum_{K,L} f_{KL}^J(g) \cdot (g^K)^{(1)}(g^L)^{(2)} \\
 &+ \sum_{K,M} f_{KM}^J(g) \cdot (g^K)^{(2)}(g^M)^{(1)} + \sum_K f_K^J(g) \cdot (g^K)^{(3)}.
 \end{aligned} \tag{2.6;3}$$

The corresponding trees are represented in the third line of Fig. 2.2. Each time we differentiate, we have to

- i) differentiate the first factor f_K^J ; i.e., we add a new branch to the root j ;
- ii) increase the derivative numbers of each of the g 's by 1; we represent this by lengthening the corresponding branch.

Each time we add a new label. *All trees which are obtained in this way are those "special" trees which have no ramifications except at the root.*

Definition 2.7. We denote by LS_q the set of *special labelled trees of order q* which have no ramifications except at the root.

Lemma 2.8 (Faà di Bruno's formula). *For $q \geq 1$ we have*

$$(f^J(g))^{(q-1)} = \sum_{u \in LS_q} \sum_{K_1, \dots, K_m} f_{K_1, \dots, K_m}^J(g) \cdot (g^{K_1})^{(\delta_1)} \dots (g^{K_m})^{(\delta_m)} \tag{2.6;q-1}$$

Here, for $u \in LS_q$, m is the number of branches leaving the root and $\delta_1, \dots, \delta_m$ are the numbers of nodes in each of these branches, such that $q = 1 + \delta_1 + \dots + \delta_m$. \square

Remark. The usual multinomial coefficients are absent here, as we use labelled trees.

The Derivatives of the Numerical Solution

It is difficult to keep a cool head when discussing the various derivatives . . .
(S. Gill 1956)

In order to generalize (2.5;1-3), we need the following definitions:

Definition 2.9. Let t be a labelled tree with root j ; we denote by

$$\Phi_j(t) = \sum_{k,l,\dots} a_{jk} a_{\dots} \dots$$

the sum over the $q-1$ remaining indices k, l, \dots (as in Definition 2.3). The summand is a product of $q-1$ a 's, where all fathers stand two by two with their sons as indices. If the set A_q is written as in (2.10), we have

$$\Phi_{j_1}(t) = \sum_{j_2, \dots, j_q} a_{t(j_2), j_2} \dots a_{t(j_q), j_q}. \quad (2.13)$$

Definition 2.10. For $t \in LT_q$ let $\gamma(t)$ be the product of $\varrho(t)$ and all orders of the trees which appear, if the roots, one after another, are removed from t . (See Fig. 2.3 or formula (2.17)).

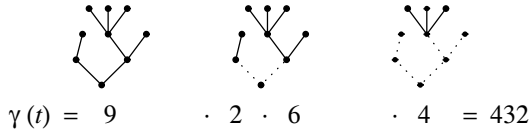


Fig. 2.3. Example for the definition of $\gamma(t)$

The above expressions are of course independent of the labellings, so $\Phi_j(t)$ as well as $\gamma(t)$ also make sense in T_q . Examples are given in Table 2.2.

Theorem 2.11. The derivatives of g_i satisfy

$$g_i^{(q)}|_{h=0} = \sum_{t \in LT_q} \gamma(t) \sum_j a_{ij} \Phi_j(t) F(t)(y_0). \quad (2.5;q)$$

The numerical solution y_1 of (2.3) satisfies

$$\begin{aligned} y_1^{(q)}|_{h=0} &= \sum_{t \in LT_q} \gamma(t) \sum_j b_j \Phi_j(t) F(t)(y_0) \\ &= \sum_{t \in T_q} \alpha(t) \gamma(t) \sum_j b_j \Phi_j(t) F(t)(y_0). \end{aligned} \quad (2.14)$$

Proof. Because of the similarity of y_1 and g_i (see (2.3)) we only have to prove the first equation. We do this by induction on q , in exactly the same way as we obtained (2.5;1-3): we first apply Leibniz' formula (2.4) to obtain

$$(g_i^J)^{(q)}|_{h=0} = q \sum_j a_{ij} (f^J(g_j))^{(q-1)}|_{y=y_0}. \quad (2.15)$$

Next we use Faà di Bruno's formula (Lemma 2.8). Finally we insert for the derivatives $(g_j^{K_s})^{(\delta_s)}$, which appear in (2.6;q-1) with $\delta_s < q$, the induction hypothesis (2.5;1) - (2.5;q-1) and rearrange the sums. This gives

$$\begin{aligned} (g_i^J)^{(q)}|_{h=0} &= q \sum_{u \in LS_q} \sum_{t_1 \in LT_{\delta_1}} \cdots \sum_{t_m \in LT_{\delta_m}} \gamma(t_1) \cdots \gamma(t_m) \cdot \\ &\quad \sum_j a_{ij} \sum_{k_1} a_{jk_1} \Phi_{k_1}(t_1) \cdots \sum_{k_m} a_{jk_m} \Phi_{k_m}(t_m) \cdot \\ &\quad \sum_{K_1, \dots, K_m} f_{K_1, \dots, K_m}^J(y_0) F^{K_1}(t_1)(y_0) \cdots F^{K_m}(t_m)(y_0). \end{aligned} \quad (2.16)$$

The main difficulty is now to understand that to each tuple

$$(u, t_1, \dots, t_m) \quad \text{with} \quad u \in LS_q, \quad t_s \in LT_{\delta_s}$$

there corresponds a labelled tree $t \in LT_q$ such that

$$\gamma(t) = q \cdot \gamma(t_1) \cdots \gamma(t_m) \quad (2.17)$$

$$F^J(t)(y) = \sum_{K_1, \dots, K_m} f_{K_1, \dots, K_m}^J(y) F^{K_1}(t_1)(y) \cdots F^{K_m}(t_m)(y) \quad (2.18)$$

$$\Phi_j(t) = \sum_{k_1, \dots, k_m} a_{jk_1} \cdots a_{jk_m} \Phi_{k_1}(t_1) \cdots \Phi_{k_m}(t_m). \quad (2.19)$$

This labelled tree t is obtained if the branches of u are replaced by the trees t_1, \dots, t_m and the corresponding labels are taken over in a natural way, i.e., in the same order (see Fig. 2.4 for some examples).

In this way, *all* trees $t \in LT_q$ appear exactly *once*. Thus (2.16) becomes (2.5;q) after inserting (2.17), (2.18) and (2.19). \square

The above construction of t can also be used for a recursive definition of trees. We first observe that the equivalence class of t (in Fig. 2.4) depends only on the equivalence classes of t_1, \dots, t_m .

Definition 2.12. We denote by

$$t = [t_1, \dots, t_m] \quad (2.20)$$

the tree, which leaves over the trees t_1, \dots, t_m when its root and the adjacent branches are chopped off (Fig. 2.5).

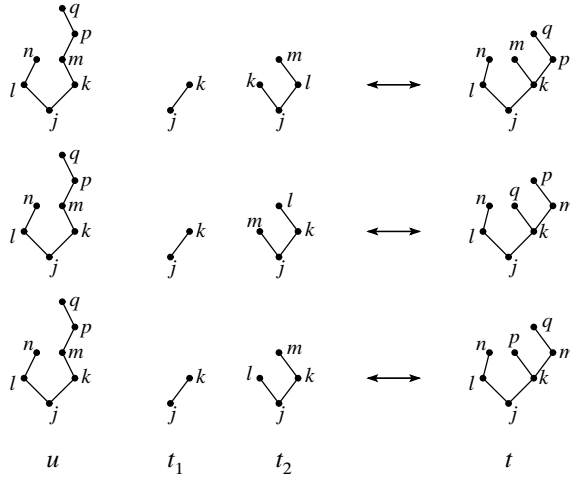


Fig. 2.4. Example for the bijection $(u, t_1, \dots, t_m) \leftrightarrow t$

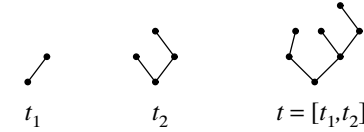


Fig. 2.5. Recursive definition of trees

With (2.20) all trees can be expressed in terms of τ ; e.g., $t_{21} = [\tau]$, $t_{31} = [\tau, \tau]$, $t_{32} = [[\tau]]$, \dots , etc.

The Order Conditions

Comparing Theorems 2.6 and 2.11 we now obtain:

Theorem 2.13. *A Runge-Kutta method (1.8) is of order p iff*

$$\sum_{j=1}^s b_j \Phi_j(t) = \frac{1}{\gamma(t)} \quad (2.21)$$

for all trees of order $\leq p$.

Proof. While the “if” part is clear from the preceding discussion, the “only if” part needs the fact that the elementary differentials for different trees are actually independent. See Exercises 3 and 4 below. \square

From Table 2.1 we then obtain the following number of order conditions (see Table 2.3). One can thus understand that the construction of higher order Runge Kutta formulas is not an easy task.

Table 2.3. Number of order conditions

order p	1	2	3	4	5	6	7	8	9	10
no. of conditions	1	2	4	8	17	37	85	200	486	1205

Example. For the tree t_{42} of Table 2.2 we have (using (1.9) for the second expression)

$$\sum_{j,k,l,m} b_j a_{jk} a_{jl} a_{km} = \sum_{j,k} b_j a_{jk} c_j c_k = \frac{1}{8},$$

which is (1.11;f). All remaining conditions of (1.11) correspond to the other trees of order ≤ 4 .

Exercises

1. Find all trees of order 6 and order 7.

Hint. Search for all representations of $p-1$ as a sum of positive integers, and then insert all known trees of lower order for each term in the sum. You may also use a computer for general p .

2. (A. Cayley 1857). Denote the number of trees of order q by a_q . Prove that

$$a_1 + a_2 x + a_3 x^2 + a_4 x^3 + \dots = (1-x)^{-a_1} (1-x^2)^{-a_2} (1-x^3)^{-a_3} \dots$$

Compare the result with Table 2.1.

3. Compute the elementary differentials of Table 2.2 for the case of the scalar non-autonomous equation (2.1), i.e., $f^1 = 1$, $f^2 = f(x, y)$. One imagines the complications met by the first authors (Kutta, Nyström, Huřa) in looking for higher order conditions. Observe also that in this case the expressions for t_{54} and t_{57} are the same, so that here Theorem 2.13 is sufficient, but not necessary for order 5.

Hint. For, say, t_{54} we have non-zero derivatives only if $K = L = 2$. Letting M and P run from 1 to 2 we then obtain

$$F^2(t) = (f_x + f f_y)(f_{yx} + f f_{yy}) f_y$$

(see also Butcher 1963a).

4. Show that for every $t \in T_q$ there is a system of differential equations such that $F^1(t)(y_0) = 1$ and $F^1(u)(y_0) = 0$ for all other trees u .

Hint. For t_{54} this system would be

$$y'_1 = y_2 y_5, \quad y'_2 = y_3, \quad y'_3 = y_4, \quad y'_4 = 1, \quad y'_5 = 1$$

with all initial values $= 0$. Understand this and the general formula

$$y'_{\text{father}} = \prod y_{\text{sons}}.$$

5. Kutta (1901) claimed that the scheme given in Table 2.4 is of order 5. Was he correct in his statement? Try to correct these values.

Result. The values for $a_{6j} (j = 1, \dots, 5)$ should read $(6, 36, 10, 8, 0)/75$; the correct values for b_j are $(23, 0, 125, 0, -81, 125)/192$ (Nyström 1925).

Table 2.4. A method of Kutta

0						
$\frac{1}{3}$	$\frac{1}{3}$					
$\frac{2}{5}$	$\frac{4}{25}$	$\frac{6}{25}$				
1	$\frac{1}{4}$	-3	$\frac{15}{4}$			
$\frac{2}{3}$	$\frac{6}{81}$	$\frac{90}{81}$	$-\frac{50}{81}$	$\frac{8}{81}$		
$\frac{4}{5}$	$\frac{7}{30}$	$\frac{18}{30}$	$-\frac{5}{30}$	$\frac{4}{30}$	0	
	$\frac{48}{192}$	0	$\frac{125}{192}$	0	$-\frac{81}{192}$	$\frac{100}{192}$

6. Verify $\sum_{\varrho(t)=p} \alpha(t) = (p-1)!$

7. Prove that a Runge-Kutta method, when applied to a linear system

$$y' = A(x)y + g(x), \quad (2.22)$$

is of order p iff

$$\sum_j b_j c_j^{q-1} = 1/q \quad \text{for } q \leq p$$

$$\sum_{j,k} b_j c_j^{q-1} a_{jk} c_k^{r-1} = 1/((q+r)r) \quad \text{for } q+r \leq p$$

$\sum_{j,k,l} b_j c_j^{q-1} a_{jk} c_k^{r-1} a_{kl} c_l^{s-1} = 1/((q+r+s)(r+s)s) \quad \text{for } q+r+s \leq p$
 ... etc (write (2.22) in autonomous form and investigate which elementary differentials vanish identically; see also Crouzeix 1975).

II.3 Error Estimation and Convergence for RK Methods

Es fehlt indessen noch der Beweis dass diese Näherungs-Verfahren convergent sind oder, was practisch wichtiger ist, es fehlt ein Kriterium, um zu ermitteln, wie klein die Schritte gemacht werden müssen, um eine vorgeschriebene Genauigkeit zu erreichen. (Runge 1905)

Since the work of Lagrange (1797) and, above all, of Cauchy, a numerically established result should be accompanied by a reliable error estimation (“... l’erreur commise sera inférieure à ...”). Lagrange gave the well-known error bounds for the Taylor polynomials and Cauchy derived bounds for the error of the Euler polygons (see Section I.7). A couple of years after the first success of the Runge-Kutta methods, Runge (1905) also required error estimates for these methods.

Rigorous Error Bounds

Runge’s device for obtaining bounds for the error in one step (“local error”) can be described in a few lines (free translation):

“For a method of order p consider the local error

$$e(h) = y(x_0 + h) - y_1 \quad (3.1)$$

and use its Taylor expansion

$$e(h) = e(0) + he'(0) + \dots + \frac{h^p}{p!}e^{(p)}(\theta h) \quad (3.2)$$

with $0 < \theta < 1$ and $e(0) = e'(0) = \dots = e^{(p)}(0) = 0$. Now compute explicitly $e^{(p)}(h)$, which will be of the form

$$e^{(p)}(h) = E_1(h) + hE_2(h), \quad (3.3)$$

where $E_1(h)$ and $E_2(h)$ contain partial derivatives of f up to order $p-1$ and p respectively. Further, because of $e^{(p)}(0) = 0$, we have $E_1(0) = 0$. Thus, if all partial derivatives of f up to order p are bounded, we have $E_1(h) = \mathcal{O}(h)$ and $E_2(h) = \mathcal{O}(1)$. So there is a constant C such that $|e^{(p)}(h)| \leq Ch$ and

$$|e(h)| \leq C \frac{h^{p+1}}{p!}. \quad (3.4)$$

A slightly different approach is adopted by Bieberbach (1923, 1. Abschn., Kap. II, §7), explained in more detail in Bieberbach (1951): we write

$$e(h) = y(x_0 + h) - y_1 = y(x_0 + h) - y_0 - h \sum_{i=1}^s b_i k_i \quad (3.5)$$

and use the Taylor expansions

$$\begin{aligned} y(x_0 + h) &= y_0 + y'(x_0)h + y''(x_0)\frac{h^2}{2!} + \dots + y^{(p+1)}(x_0 + \theta h)\frac{h^{p+1}}{(p+1)!} \\ k_i(h) &= k_i(0) + k'_i(0)h + \dots + k_i^{(p)}(\theta_i h)\frac{h^p}{p!}, \end{aligned} \quad (3.6)$$

where, for vector valued functions, the formula is valid componentwise with possibly different θ 's. The first terms in the h expansion of (3.5) vanish because of the order conditions. Thus we obtain

Theorem 3.1. *If the Runge-Kutta method (1.8) is of order p and if all partial derivatives of $f(x, y)$ up to order p exist (and are continuous), then the local error of (1.8) admits the rigorous bound*

$$\begin{aligned} \|y(x_0 + h) - y_1\| &\leq h^{p+1} \left(\frac{1}{(p+1)!} \max_{t \in [0,1]} \|y^{(p+1)}(x_0 + th)\| \right. \\ &\quad \left. + \frac{1}{p!} \sum_{i=1}^s |b_i| \max_{t \in [0,1]} \|k_i^{(p)}(th)\| \right) \end{aligned} \quad (3.7)$$

and hence also

$$\|y(x_0 + h) - y_1\| \leq Ch^{p+1}. \quad (3.8)$$

□

Let us demonstrate this result on Runge's first method (1.4), which is of order $p = 2$, applied to a scalar differential equation. Differentiating (1.1) we obtain

$$y^{(3)}(x) = \left(f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_y(f_x + f_yf) \right) (x, y(x)) \quad (3.9)$$

while the second derivative of $k_2(h) = f(x_0 + \frac{h}{2}, y_0 + \frac{h}{2}f_0)$ is given by

$$k_2^{(2)}(h) = \frac{1}{4} \left(f_{xx}(x_0 + \frac{h}{2}, y_0 + \frac{h}{2}f_0) + 2f_{xy}(\dots)f_0 + f_{yy}(\dots)f_0^2 \right) \quad (3.10)$$

(f_0 stands for $f(x_0, y_0)$). Under the assumptions of Theorem 3.1 we see that the expressions (3.9) and (3.10) are bounded by a constant independent of h , which gives (3.8).

The Principal Error Term

For higher order methods rigorous error bounds, like (3.7), become very unpractical. It is therefore much more realistic to consider the first non-zero term in the Taylor expansion of the error. For autonomous systems of equations (2.2), the error term is best obtained by subtracting the Taylor series and using (2.14) and (2.7;q).

Theorem 3.2. *If the Runge-Kutta method is of order p and if f is $(p+1)$ -times continuously differentiable, we have*

$$y^J(x_0 + h) - y_1^J = \frac{h^{p+1}}{(p+1)!} \sum_{t \in T_{p+1}} \alpha(t) e(t) F^J(t)(y_0) + \mathcal{O}(h^{p+2}) \quad (3.11)$$

where

$$e(t) = 1 - \gamma(t) \sum_{j=1}^s b_j \Phi_j(t). \quad (3.12)$$

□

$\gamma(t)$ and $\Phi_j(t)$ are given in Definitions 2.9 and 2.10; see also formulas (2.17) and (2.19). The expressions $e(t)$ are called the *error coefficients*.

Example 3.3. For the two-parameter family of 4th order RK methods (1.17) the error coefficients for the 9 trees of Table 2.2 are ($c_2 = u$, $c_3 = v$):

$$\begin{aligned} e(t_{51}) &= -\frac{1}{4} + \frac{5}{12}(u+v) - \frac{5}{6}uv, & e(t_{52}) &= \frac{5}{12}v - \frac{1}{4}, \\ e(t_{53}) &= \frac{5}{8}u - \frac{1}{4}, & e(t_{54}) &= -\frac{1}{4}, \\ e(t_{55}) &= 1 - \frac{5(b_4 + b_3(3-4v)^2)}{144b_3b_4(1-v)^2}, & & \\ e(t_{56}) &= -4e(t_{51}), & e(t_{57}) &= -4e(t_{52}), \\ e(t_{58}) &= -4e(t_{53}), & e(t_{59}) &= -4e(t_{54}). \end{aligned} \quad (3.13)$$

Proof. The last four formulas follow from (1.12). $e(t_{59})$ is trivial, $e(t_{58})$ and $e(t_{57})$ follow from (1.11h). Further

$$e(t_{51}) = 5 \int_0^1 t(t-1)(t-u)(t-v) dt$$

expresses the quadrature error. For $e(t_{55})$ one best introduces $c'_i = \sum_j a_{ij}c_j$ such that $e(t_{55}) = 1 - 20 \sum_i b_i c'_i c'_i$. Then from (1.1d,f) one obtains

$$c'_1 = c'_2 = 0, \quad b_3 c'_3 = \frac{1}{24(1-v)}, \quad b_4 c'_4 = \frac{3-4v}{24(1-v)}. \quad \square$$

For the classical 4th order method (Table 1.2a) these error coefficients are given by Kutta (1901), p. 448 (see also Lotkin 1951) as follows

$$\left(-\frac{1}{24}, -\frac{1}{24}, \frac{1}{16}, -\frac{1}{4}, -\frac{2}{3}, \frac{1}{6}, \frac{1}{6}, -\frac{1}{4}, 1\right)$$

Kutta remarked that for the second method (Table 1.2b) (“Als besser noch erweist sich . . .”) the error coefficients become

$$\left(-\frac{1}{54}, \frac{1}{36}, -\frac{1}{24}, -\frac{1}{4}, -\frac{1}{9}, \frac{2}{27}, -\frac{1}{9}, \frac{1}{6}, 1\right)$$

which, with the exception of the 4th and 9th term, are all smaller than for the above method. A tedious calculation was undertaken by Ralston (1962) (and by many others) to determine optimal coefficients of (1.17). For solutions which minimize the constants (3.13), see Exercise 3 below.

Estimation of the Global Error

(P. Henrici 1983)

The global error is the error of the computed solution after *several* steps. Suppose that we have a one-step method which, given an initial value (x_0, y_0) and a step size h , computes a numerical solution y_1 approximating $y(x_0 + h)$. We shall denote this process by Henrici’s notation

$$y_1 = y_0 + h\Phi(x_0, y_0, h) \quad (3.14)$$

and call Φ the *increment function* of the method.

The numerical solution for a point $X > x_0$ is then obtained by a step-by-step procedure

$$y_{i+1} = y_i + h_i\Phi(x_i, y_i, h_i), \quad h_i = x_{i+1} - x_i, \quad x_N = X \quad (3.15)$$

and our task is to estimate the *global error*

$$E = y(X) - y_N. \quad (3.16)$$

This estimate is found in a simple way, very similar to Cauchy’s convergence proof for Theorem 7.3 of Chapter I: *the local errors are transported to the final point x_N and then added up*. This “error transport” can be done in two different ways:

a) either along the exact solution curves (see Fig. 3.1); this method can yield sharp results when sharp estimates of error propagation for the exact solutions are known, e.g., from Theorem 10.6 of Chapter I based on the logarithmic norm $\mu(\partial f/\partial y)$.

b) or along $N - i$ steps of the numerical method (see Fig. 3.2); this is the method used in the proofs of Cauchy (1824) and Runge (1905), it generalizes easily to multistep methods (see Chapter III) and will be an important tool for the existence of asymptotic expansions (see II.8).

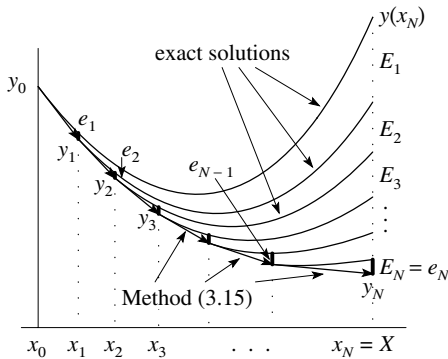


Fig. 3.1. Global error estimation, method (a)

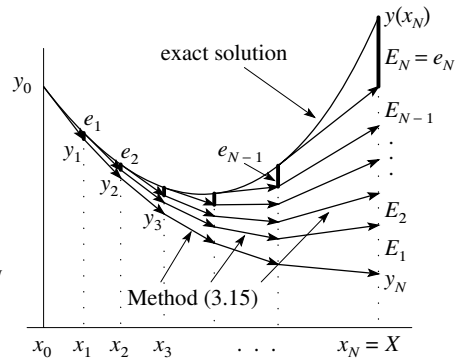


Fig. 3.2. Global error estimation, method (b)

In both cases we first estimate the local errors e_i with the help of Theorem 3.1 to obtain

$$\|e_i\| \leq C \cdot h_{i-1}^{p+1}. \quad (3.17)$$

Warning. The e_i of Fig.3.1 and Fig. 3.2, for $i \neq 1$, are *not* the same, but they allow similar estimates.

We then estimate the transported errors E_i : for method (a) we use the known results from Chapter I, especially Theorem I.10.6, Theorem I.10.2, or formula (I.7.17). The result is

Theorem 3.4. Let U be a neighbourhood of $\{(x, y(x)) | x_0 \leq x \leq X\}$ where $y(x)$ is the exact solution of (1.1). Suppose that in U

$$\left\| \frac{\partial f}{\partial y} \right\| \leq L \quad \text{or} \quad \mu \left(\frac{\partial f}{\partial y} \right) \leq L, \quad (3.18)$$

and that the local error estimates (3.17) are valid in U . Then the global error (3.16) can be estimated by

$$\|E\| \leq h^p \frac{C'}{L} \left(\exp(L(X - x_0)) - 1 \right) \quad (3.19)$$

where $h = \max h_i$,

$$C' = \begin{cases} C & L \geq 0 \\ C \exp(-Lh) & L < 0, \end{cases}$$

and h is small enough for the numerical solution to remain in U .

Remark. For $L \rightarrow 0$ the estimate (3.19) tends to $h^p C (x_N - x_0)$.

Proof. From Theorem I.10.2 (with $\varepsilon = 0$) or Theorem I.10.6 (with $\delta = 0$) we obtain

$$\|E_i\| \leq \exp(L(x_N - x_i)) \|e_i\|. \quad (3.20)$$

We then insert this together with (3.17) into

$$\|E\| \leq \sum_{i=1}^N \|E_i\|.$$

Using $h_{i-1}^{p+1} \leq h^p \cdot h_{i-1}$ this leads to

$$\|E\| \leq h^p C \left(h_0 \exp(L(x_N - x_1)) + h_1 \exp(L(x_N - x_2)) + \dots \right).$$

The expression in large brackets can be bounded by

$$\int_{x_0}^{x_N} \exp(L(x_N - x)) dx \quad \text{for} \quad L \geq 0 \quad (3.21)$$

$$\int_{x_0}^{x_N} \exp(L(x_N - h - x)) dx \quad \text{for} \quad L < 0 \quad (3.22)$$

(see Fig. 3.3). This gives (3.19). \square

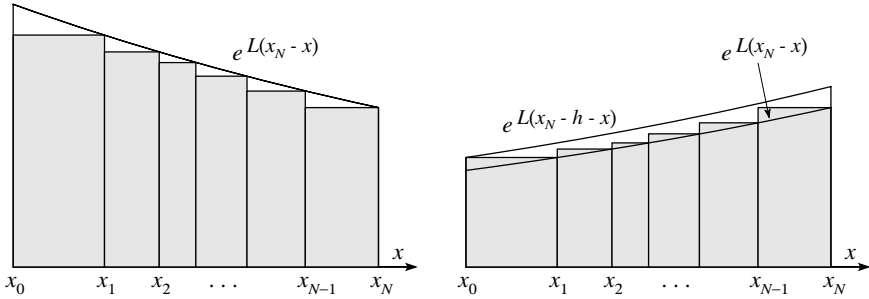


Fig. 3.3. Estimation of Riemann sums

For the second method (b) we need an estimate for $\|z_{i+1} - y_{i+1}\|$ in terms of $\|z_i - y_i\|$, where, besides (3.15),

$$z_{i+1} = z_i + h_i \Phi(x_i, z_i, h_i)$$

is a second pair of numerical solutions. For RK-methods z_{i+1} is defined by

$$\begin{aligned} \ell_1 &= f(x_i, z_i), \\ \ell_2 &= f(x_i + c_2 h_i, z_i + h_i a_{21} \ell_1), \quad \text{etc.} \end{aligned}$$

We now subtract formulas (1.8) from this and obtain

$$\begin{aligned}\|\ell_1 - k_1\| &\leq L\|z_i - y_i\|, \\ \|\ell_2 - k_2\| &\leq L(1 + |a_{21}|h_iL)\|z_i - y_i\|, \quad \text{etc.}\end{aligned}$$

This leads to the following

Lemma 3.5. *Let L be a Lipschitz constant for f and let $h_i \leq h$. Then the increment function Φ of method (1.8) satisfies*

$$\|\Phi(x_i, z_i, h_i) - \Phi(x_i, y_i, h_i)\| \leq \Lambda\|z_i - y_i\| \quad (3.23)$$

where

$$\Lambda = L\left(\sum_i |b_i| + hL \sum_{i,j} |b_i a_{ij}| + h^2 L^2 \sum_{i,j,k} |b_i a_{ij} a_{jk}| + \dots\right). \quad (3.24)$$

□

From (3.23) we obtain

$$\|z_{i+1} - y_{i+1}\| \leq (1 + h_i \Lambda)\|z_i - y_i\| \leq \exp(h_i \Lambda)\|z_i - y_i\| \quad (3.25)$$

and for the errors in Fig. 3.2,

$$\|E_i\| \leq \exp(\Lambda(x_N - x_i))\|e_i\| \quad (3.26)$$

instead of (3.20). The same proof as for Theorem 3.4 now gives us

Theorem 3.6. *Suppose that the local error satisfies, for initial values on the exact solution,*

$$\|y(x+h) - y(x) - h\Phi(x, y(x), h)\| \leq Ch^{p+1}, \quad (3.27)$$

and suppose that in a neighbourhood of the solution the increment function Φ satisfies

$$\|\Phi(x, z, h) - \Phi(x, y, h)\| \leq \Lambda\|z - y\|. \quad (3.28)$$

Then the global error (3.16) can be estimated by

$$\|E\| \leq h^p \frac{C}{\Lambda} \left(\exp(\Lambda(x_N - x_0)) - 1 \right) \quad (3.29)$$

where $h = \max h_i$.

□

Exercises

1. (Runge 1905). Show that for explicit Runge Kutta methods with $b_i \geq 0$, $a_{ij} \geq 0$ (all i, j) of order s the Lipschitz constant Λ for Φ satisfies

$$1 + h\Lambda < \exp(hL)$$

and that (3.29) is valid with Λ replaced by L .

2. Show that $e(t_{55})$ of (3.13) becomes

$$e(t_{55}) = 1 - 5 \frac{(4v^2 - 15v + 9) - u(6v^2 - 42v + 27) - u^2(26v - 18)}{12(1 - 2u)(6uv - 4(u + v) + 3)}$$

after inserting (1.17).

3. Determine u and v in (1.17) such that in (3.13)

$$\begin{aligned} \text{a) } \max_{i=5,6,7,8} |e(t_{5i})| &= \min & \text{b) } \sum_{i=1}^9 |e(t_{5i})| &= \min \\ \text{c) } \max_{i=5,6,7,8} \alpha(t_{5i}) |e(t_{5i})| &= \min & \text{d) } \sum_{i=1}^9 \alpha(t_{5i}) |e(t_{5i})| &= \min \end{aligned}$$

Results.

$$\begin{aligned} \text{a) } u &= 0.3587, \quad v = 0.6346, \quad \min = 0.1033; \\ \text{b) } u &= 0.3995, \quad v = 0.6, \quad \min = 1.55; \\ \text{c) } u &= 0.3501, \quad v = 0.5839, \quad \min = 0.1248; \\ \text{d) } u &= 0.3716, \quad v = 0.6, \quad \min = 2.53. \end{aligned}$$

Such optimal formulas were first studied by Ralston (1962), Hull & Johnston (1964), and Hull (1967).

4. Apply an explicit Runge-Kutta method to the problem $y' = f(x, y)$, $y(0) = 0$, where

$$f(x, y) = \begin{cases} \frac{\lambda}{x} y + g(x) & \text{if } x > 0 \\ (1 - \lambda)^{-1} g(0) & \text{if } x = 0, \end{cases}$$

$\lambda \leq 0$ and $g(x)$ is sufficiently differentiable (see Exercise 10 of Section I.5).

- a) Show that the error after the first step is given by

$$y(h) - y_1 = C_2 h^2 g'(0) + \mathcal{O}(h^3)$$

where C_2 is a constant depending on λ and on the coefficients of the method. Also for high order methods we have in general $C_2 \neq 0$.

- b) Compute C_2 for the classical 4th order method (Table 1.2).

II.4 Practical Error Estimation and Step Size Selection

Ich glaube indessen, dass ein practischer Rechner sich meistens mit der geringeren Sicherheit begnügen wird, die er aus der Uebereinstimmung seiner Resultate für grössere und kleinere Schritte gewinnt. (C. Runge 1895)

Even the simplified error estimates of Section II.3, which are content with the leading error term, are of little practical interest, because they require the computation and majorization of several partial derivatives of high orders. But the main advantage of Runge-Kutta methods, compared with Taylor series, is precisely that the computation of derivatives should be no longer necessary. However, since practical error estimates are necessary (on the one hand to ensure that the step sizes h_i are chosen sufficiently small to yield the required precision of the computed results, and on the other hand to ensure that the step sizes are sufficiently large to avoid unnecessary computational work), we shall now discuss alternative methods for error estimates.

The oldest device, used by Runge in his numerical examples, is to repeat the computations with *halved* step sizes and to compare the results: those digits which haven't changed are assumed to be correct (“... woraus ich schliessen zu dürfen glaube ...”).

Richardson Extrapolation

... its usefulness for practical computations can hardly be overestimated. (G. Birkhoff & G.C. Rota)

The idea of Richardson, announced in his classical paper Richardson (1910) which treats mainly partial differential equations, and explained in full detail in Richardson (1927), is to use more carefully the known behaviour of the error as a function of h .

Suppose that, with a given initial value (x_0, y_0) and step size h , we compute *two* steps, using a fixed Runge-Kutta method of order p , and obtain the numerical results y_1 and y_2 . We then compute, starting from (x_0, y_0) , *one big step* with step size $2h$ to obtain the solution w . The error of y_1 is known to be (Theorem 3.2)

$$e_1 = y(x_0 + h) - y_1 = C \cdot h^{p+1} + \mathcal{O}(h^{p+2}) \quad (4.1)$$

where C contains the error coefficients of the method and the elementary differentials $F^J(t)(y_0)$ of order $p+1$. The error of y_2 is composed of two parts: the

transported error of the first step, which is

$$\left(I + h \frac{\partial f}{\partial y} + \mathcal{O}(h^2)\right)e_1,$$

and the local error of the second step, which is the same as (4.1), but with the elementary differentials evaluated at $y_1 = y_0 + \mathcal{O}(h)$. Thus we obtain

$$\begin{aligned} e_2 = y(x_0 + 2h) - y_2 &= (I + \mathcal{O}(h))Ch^{p+1} + (C + \mathcal{O}(h))h^{p+1} + \mathcal{O}(h^{p+2}) \\ &= 2Ch^{p+1} + \mathcal{O}(h^{p+2}). \end{aligned} \quad (4.2)$$

Similarly to (4.1), we have for the big step

$$y(x_0 + 2h) - w = C(2h)^{p+1} + \mathcal{O}(h^{p+2}). \quad (4.3)$$

Neglecting the terms $\mathcal{O}(h^{p+2})$, formulas (4.2) and (4.3) allow us to eliminate the unknown constant C and to “extrapolate” a better value \hat{y}_2 for $y(x_0 + 2h)$, for which we obtain:

Theorem 4.1. *Suppose that y_2 is the numerical result of two steps with step size h of a Runge-Kutta method of order p , and w is the result of one big step with step size $2h$. Then the error of y_2 can be extrapolated as*

$$y(x_0 + 2h) - y_2 = \frac{y_2 - w}{2^p - 1} + \mathcal{O}(h^{p+2}) \quad (4.4)$$

and

$$\hat{y}_2 = y_2 + \frac{y_2 - w}{2^p - 1} \quad (4.5)$$

is an approximation of order $p + 1$ to $y(x_0 + 2h)$. □

Formula (4.4) is a very simple device to estimate the error of y_2 and formula (4.5) allows one to increase the precision by one additional order (“... The better theory of the following sections is complicated, and tends thereby to suggest that the practice may also be complicated; whereas it is really simple.” Richardson).

Embedded Runge-Kutta Formulas

Scraton is right in his criticism of Merson’s process, although Merson did not claim as much for his process as some people expect. (R. England 1969)

The idea is, rather than using Richardson extrapolation, to construct Runge-Kutta formulas which themselves contain, besides the numerical approximation y_1 , a second approximation \hat{y}_1 . The difference then yields an estimate of the local error for the less precise result and can be used for step size control (see below). Since

it is at our disposal at every step, this gives more flexibility to the code and makes step rejections less expensive.

We consider two Runge-Kutta methods (one for y_1 and one for \hat{y}_1) such that both use the *same* function values. We thus have to find a scheme of coefficients (see (1.8')),

$$\begin{array}{c|cccc}
 0 & & & & \\
 c_2 & a_{21} & & & \\
 c_3 & a_{32} & a_{32} & & \\
 \vdots & \vdots & \ddots & & \\
 c_s & a_{s1} & a_{s2} & \dots & a_{s,s-1} \\
 \hline
 & b_1 & b_2 & \dots & b_{s-1} & b_s \\
 & \hat{b}_1 & \hat{b}_2 & \dots & \hat{b}_{s-1} & \hat{b}_s
 \end{array} \tag{4.6}$$

such that

$$y_1 = y_0 + h(b_1 k_1 + \dots + b_s k_s) \tag{4.7}$$

is of order p , and

$$\hat{y}_1 = y_0 + h(\hat{b}_1 k_1 + \dots + \hat{b}_s k_s) \tag{4.7'}$$

is of order \hat{p} (usually $\hat{p} = p - 1$ or $\hat{p} = p + 1$). The approximation y_1 is used to continue the integration.

From Theorem 2.13, we have to satisfy the conditions

$$\sum_{j=1}^s b_j \Phi_j(t) = \frac{1}{\gamma(t)} \quad \text{for all trees of order } \leq p, \tag{4.8}$$

$$\sum_{j=1}^s \hat{b}_j \Phi_j(t) = \frac{1}{\gamma(t)} \quad \text{for all trees of order } \leq \hat{p}. \tag{4.8'}$$

The first methods of this type were proposed by Merson (1957), Ceschino (1962), and Zonneveld (1963). Those of Merson and Zonneveld are given in Tables 4.1 and 4.2. Here, “name $p(\hat{p})$ ” means that the order of y_1 is p and the order of the error estimator \hat{y}_1 is \hat{p} . Merson’s \hat{y}_1 is of order 5 only for *linear* equations with constant coefficients; for nonlinear problems it is of order 3. This method works quite well and has been used very often, especially by NAG users. Further embedded methods were then derived by Sarafyan (1966), England (1969), and Fehlberg (1964, 1968, 1969). Let us start with the construction of some low order embedded methods.

Methods of order 3(2). It is a simple task to construct embedded formulas of order 3(2) with $s = 3$ stages. Just take a 3-stage method of order 3 (Exercise II.1.4) and put $\hat{b}_3 = 0$, $\hat{b}_2 = 1/2c_2$, $\hat{b}_1 = 1 - 1/2c_2$.

Table 4.1. Merson 4(“5”)

0					
$\frac{1}{3}$	$\frac{1}{3}$				
$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$			
$\frac{1}{2}$	$\frac{1}{8}$	0	$\frac{3}{8}$		
1	$\frac{1}{2}$	0	$-\frac{3}{2}$	2	
y_1	$\frac{1}{6}$	0	0	$\frac{2}{3}$	$\frac{1}{6}$
\hat{y}_1	$\frac{1}{10}$	0	$\frac{3}{10}$	$\frac{2}{5}$	$\frac{1}{5}$

Table 4.2. Zonneveld 4(3)

0					
$\frac{1}{2}$	$\frac{1}{2}$				
$\frac{1}{2}$	0	$\frac{1}{2}$			
1	0	0	1		
$\frac{3}{4}$	$\frac{5}{32}$	$\frac{7}{32}$	$\frac{13}{32}$	$-\frac{1}{32}$	
y_1	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$	
\hat{y}_1	$-\frac{1}{2}$	$\frac{7}{3}$	$\frac{7}{3}$	$\frac{13}{6}$	$-\frac{16}{3}$

Methods of order 4(3). With $s = 4$ it is impossible to find a pair of order 4(3) (see Exercise 2). The idea is to add y_1 as 5th stage of the process (i.e., $a_{5i} = b_i$ for $i = 1, \dots, 4$) and to search for a third order method which uses all five function values. Whenever the step is accepted this represents no extra work, because $f(x_0 + h, y_1)$ has to be computed anyway for the following step. This idea is called FSAL (First Same As Last). Then the order conditions (4.8') with $\hat{p} = 3$ represent 4 linear equations for the five unknowns $\hat{b}_1, \dots, \hat{b}_5$. One can arbitrarily fix $\hat{b}_5 \neq 0$ and solve the system for the remaining parameters. With \hat{b}_5 chosen such that $\hat{b}_4 = 0$ the result is

$$\begin{aligned} \hat{b}_1 &= 2b_1 - 1/6, & \hat{b}_2 &= 2(1 - c_2)b_2, \\ \hat{b}_3 &= 2(1 - c_3)b_3, & \hat{b}_4 &= 0, & \hat{b}_5 &= 1/6. \end{aligned} \quad (4.9)$$

Automatic Step Size Control

D'ordinaire, on se contente de multiplier ou de diviser par 2 la valeur du pas ... (Ceschino 1961)

We now want to write a code which automatically adjusts the step size in order to achieve a prescribed tolerance of the local error.

Whenever a starting step size h has been chosen, the program computes two approximations to the solution, y_1 and \hat{y}_1 . Then an estimate of the error for the less precise result is $y_1 - \hat{y}_1$. We want this error to satisfy componentwise

$$|y_{1i} - \hat{y}_{1i}| \leq sc_i, \quad sc_i = Atol_i + \max(|y_{0i}|, |y_{1i}|) \cdot Rtol_i \quad (4.10)$$

where $Atol_i$ and $Rtol_i$ are the desired tolerances prescribed by the user (relative errors are considered for $Atol_i = 0$, absolute errors for $Rtol_i = 0$; usually both

tolerances are different from zero; they may depend on the component of the solution). As a measure of the error we take

$$err = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{y_{1i} - \hat{y}_{1i}}{sc_i} \right)^2}; \quad (4.11)$$

other norms, such as the max norm, are also of frequent use. Then err is compared to 1 in order to find an optimal step size. From the error behaviour $err \approx C \cdot h^{q+1}$ and from $1 \approx C \cdot h_{opt}^{q+1}$ (where $q = \min(p, \hat{p})$) the optimal step size is obtained as (“... le procédé connu”, Ceschino 1961)

$$h_{opt} = h \cdot (1/err)^{1/(q+1)}. \quad (4.12)$$

Some care is now necessary for a good code: we multiply (4.12) by a safety factor fac , usually $fac = 0.8, 0.9, (0.25)^{1/(q+1)}$, or $(0.38)^{1/(q+1)}$, so that the error will be acceptable the next time with high probability. Further, h is not allowed to increase nor to decrease too fast. For example, we may put

$$h_{new} = h \cdot \min(facmax, \max(facmin, fac \cdot (1/err)^{1/(q+1)})) \quad (4.13)$$

for the new step size. Then, if $err \leq 1$, the computed step is *accepted* and the solution is advanced with y_1 and a new step is tried with h_{new} as step size. Else, the step is *rejected* and the computations are repeated with the new step size h_{new} . The maximal step size increase $facmax$, usually chosen between 1.5 and 5, prevents the code from too large step increases and contributes to its safety. It is clear that, when chosen too small, it may also unnecessarily increase the computational work. It is also advisable to put $facmax = 1$ in the steps right after a step-rejection (Shampine & Watts 1979).

Whenever y_1 is of lower order than \hat{y}_1 , then the difference $y_1 - \hat{y}_1$ is (at least asymptotically) an estimate of the local error and the above algorithm keeps this estimate below the given tolerance. But isn't it more natural to continue the integration with the higher order approximation? Then the concept of “error estimation” is abandoned and the difference $y_1 - \hat{y}_1$ is only used for the purpose of step size selection. This is justified by the fact that, due to unknown stability and instability properties of the differential system, the local errors have in general very little in common with the global errors. The procedure of continuing the integration with the higher order result is called “local extrapolation”.

A modification of the above procedure (PI step size control), which is particularly interesting when applied to mildly stiff problems, is described in Section IV.2 (Volume II).

Starting Step Size

If anything has been made foolproof, a better fool will be developed.
(Heard from Dr. Pirkle, Baden)

For many years, the starting step size had to be supplied to a code. Users were assumed to have a rough idea of a good step size from mathematical knowledge or previous experience. Anyhow, a bad starting choice for h was quickly repaired by the step size control. Nevertheless, when this happens too often and when the choices are too bad, much computing time can be wasted. Therefore, several people (e.g., Watts 1983, Hindmarsh 1980) developed ideas to let the computer do this choice. We take up an idea of Gladwell, Shampine & Brankin (1987) which is based on the hypothesis that

$$\text{local error} \approx Ch^{p+1}y^{(p+1)}(x_0).$$

Since $y^{(p+1)}(x_0)$ is unknown we shall replace it by approximations of the first and second derivative of the solution. The resulting algorithm is the following one:

- a) Do one function evaluation $f(x_0, y_0)$ at the initial point. It is in any case needed for the first RK step. Then put $d_0 = \|y_0\|$ and $d_1 = \|f(x_0, y_0)\|$, where the norm is that of (4.11) with $sc_i = Atol_i + |y_{0i}| \cdot Rtol_i$.
- b) As a first guess for the step size let

$$h_0 = 0.01 \cdot (d_0/d_1)$$

so that the increment of an explicit Euler step is small compared to the size of the initial value. If either d_0 or d_1 is smaller than 10^{-5} we put $h_0 = 10^{-6}$.

- c) Perform one explicit Euler step, $y_1 = y_0 + h_0 f(x_0, y_0)$, and compute $f(x_0 + h_0, y_1)$.
- d) Compute $d_2 = \|f(x_0 + h_0, y_1) - f(x_0, y_0)\|/h_0$ as an estimate of the second derivative of the solution; the norm being the same as in (a).
- e) Compute a step size h_1 from the relation

$$h_1^{p+1} \cdot \max(d_1, d_2) = 0.01.$$

If $\max(d_1, d_2) \leq 10^{-15}$ we put $h_1 = \max(10^{-6}, h_0 \cdot 10^{-3})$.

- f) Finally we propose as starting step size

$$h = \min(100 \cdot h_0, h_1). \quad (4.14)$$

An algorithm like the one above, or a similar one, usually gives a good guess for the initial step size (or at least avoids a very bad choice). Sometimes, more information about h is known, e.g., from previous experience or computations of similar problems.

Numerical Experiments

As a representative of 4-stage 4th order methods we consider the “3/8 Rule” of Table 1.2. We equipped it with the embedded formula (4.9) of order 3.

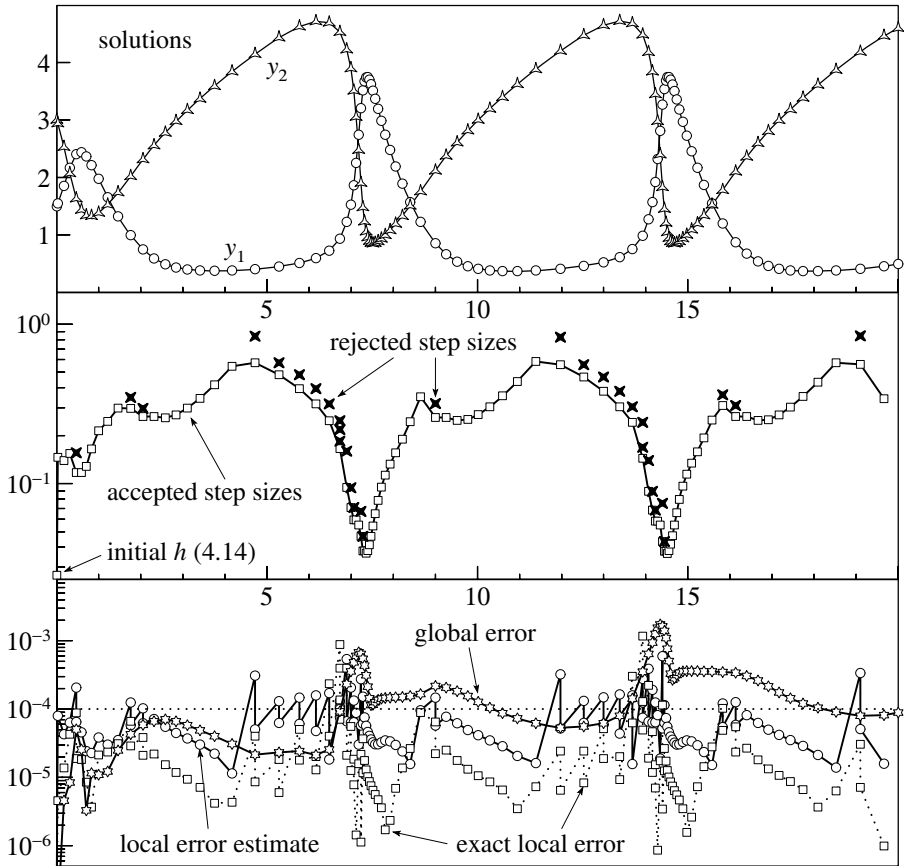


Fig. 4.1. Step size control, $Rtol = Atol = 10^{-4}$, 96 steps + 32 rejected

Step control mechanism. Fig. 4.1 presents the results of the step control mechanism (4.13) described above. As an example we choose the Brusselator (see Section I.16).

$$\begin{aligned} y_1' &= 1 + y_1^2 y_2 - 4y_1 \\ y_2' &= 3y_1 - y_1^2 y_2 \end{aligned} \quad (4.15)$$

with initial values $y_1(0) = 1.5$, $y_2(0) = 3$, integration interval $0 \leq x \leq 20$ and $Atol = Rtol = 10^{-4}$. The following results are plotted in this figure:

- i) At the top, the solutions $y_1(x)$ and $y_2(x)$ with all accepted integration steps;
- ii) then all step sizes used; the accepted ones are connected by a polygon; the rejected ones are indicated by \times ;
- iii) the third graph shows the local error estimate err , the exact local error and the global error; the desired tolerance is indicated by a broken horizontal line.

It can be seen that, due to the instabilities of the solutions with respect to the initial values, quite large global errors occur during the integration with small local tolerances everywhere. Further many step rejections can be observed in regions where the step size has to be decreased. This cannot easily be prevented, because right after an accepted step, the step size proposed by formula (4.13) is (apart from the safety factor) always increasing.

Numerical comparison. We are now curious to see the behaviour of the variable step size code, when compared to a fixed step size implementation. We applied both implementations to the Brusselator problem (4.15) with the initial values used there. The tolerances ($Atol = Rtol$) are chosen between 10^{-2} and 10^{-10} with ratio $\sqrt[3]{10}$. The results are then plotted in Fig. 4.2. There, the abscissa is the global error at the endpoint of integration (the “precision”), and the ordinate is the number of function evaluations (the “work”). We observe that for this problem the variable step size code is about twice as fast as the fixed step size code. There are, of course, problems (such as equation (0.1)) where variable step sizes are *much* more important than here.

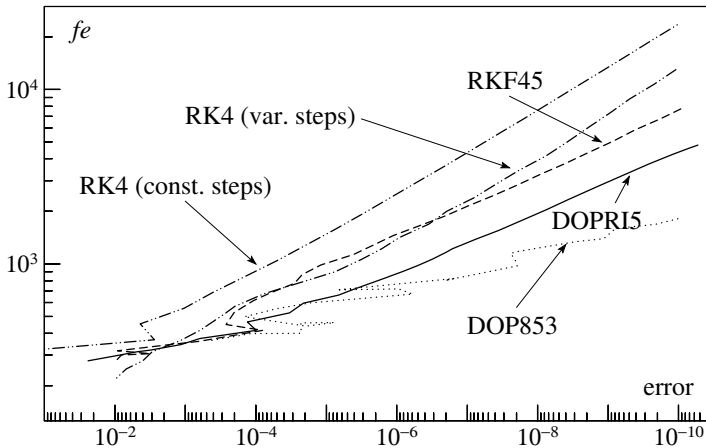


Fig. 4.2. Precision-Work diagram

In this comparison we have included some higher order methods, which will be discussed in Section II.5. The code RKF45 (written by H.A. Watts and L.F. Shampine) is based on an embedded method of order 5(4) due to Fehlberg. The codes DOPRI5 (order 5(4)) and DOP853 (order 8(5,3)) are based on methods of

Dormand & Prince. They will be discussed in the following section. It can clearly be seen that higher order methods are, especially for higher precision, more efficient than lower order methods. We shall also understand why the 5th order method of Dormand & Prince is clearly superior to RKF45.

Exercises

1. Show that Runge's method (1.4) can be interpreted as two Euler steps (with step size $h/2$), followed by a Richardson extrapolation.
2. Prove that no 4-stage Runge-Kutta method of order 4 admits an embedded formula of order 3.

Hint. Replace d_j by $\hat{b}_j - b_j$ in the proof of Lemma 1.4 and deduce that $\hat{b}_j = b_j$ for all j , which is a contradiction.

3. Show that the step size strategy (4.13) is invariant with respect to a rescaling of the independent variable. This means that it produces equivalent step size sequences when applied to the two problems

$$\begin{aligned} y' &= f(x, y), & y(0) &= y_0, & y(x_{\text{end}}) &=? \\ z' &= \sigma \cdot f(\sigma t, z), & z(0) &= y_0, & z(x_{\text{end}}/\sigma) &=? \end{aligned}$$

with initial step sizes h_0 and h_0/σ , respectively.

Remark. This is no longer the case if one replaces err in (4.13) by err/h and q by $q - 1$ ("error per unit step").

II.5 Explicit Runge-Kutta Methods of Higher Order

Gehen wir endlich zu Näherungen von der fünften Ordnung über,
so werden die Verhältnisse etwas andere. (W. Kutta 1901)

This section describes the construction of Runge-Kutta methods of higher orders, particularly of orders $p = 5$ and $p = 8$. As can be seen from Table 2.3, the complexity and number of the order conditions to be solved increases rapidly with p . An increasingly skilful use of simplifying assumptions will be the main tool for this task.

The Butcher Barriers

For methods of order 5 there are 17 order conditions to be satisfied (see Table 2.2). If we choose $s = 5$ we have 15 free parameters. Already Kutta raised the question whether there might nevertheless exist a solution (“Nun wäre es zwar möglich . . .”), but he had no hope for this and turned straight away to the case $s = 6$ (see II.2, Exercise 5). Kutta’s question remained open for more than 60 years and was answered around 1963 by three authors independently (Ceschino & Kuntzmann 1963, p. 89, Shanks 1966, Butcher 1964b, 1965b). Butcher’s work is the farthest reaching and we shall mainly follow his ideas in the following:

Theorem 5.1. *For $p \geq 5$ no explicit Runge-Kutta method exists of order p with $s = p$ stages.*

Proof. We first treat the case $s = p = 5$: define the matrices U and V by

$$U = \begin{pmatrix} \sum_i b_i a_{i2} & \sum_i b_i a_{i3} & \sum_i b_i a_{i4} \\ \sum_i b_i a_{i2} c_2 & \sum_i b_i a_{i3} c_3 & \sum_i b_i a_{i4} c_4 \\ g_2 & g_3 & g_4 \end{pmatrix}, \quad V = \begin{pmatrix} c_2 & c_2^2 & \sum_j a_{2j} c_j - c_2^2/2 \\ c_3 & c_3^2 & \sum_j a_{3j} c_j - c_3^2/2 \\ c_4 & c_4^2 & \sum_j a_{4j} c_j - c_4^2/2 \end{pmatrix} \quad (5.1)$$

where

$$g_k = \sum_{i,j} b_i a_{ij} a_{jk} - \frac{1}{2} \sum_i b_i a_{ik} (1 - c_k). \quad (5.2)$$

Then the order conditions for order 5 imply

$$UV = \begin{pmatrix} 1/6 & 1/12 & 0 \\ 1/12 & 1/20 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \quad (5.3)$$

Lemma 1.5 gives $g_4 = 0$ and consequently $c_4 = 1$ as in Lemma 1.4. Next we put in (5.1)

$$g_j = \left(\sum_i b_i a_{ij} - b_j(1 - c_j) \right) (c_j - c_5). \quad (5.4)$$

Again it can be verified by trivial computations that UV is the same as above. This time it follows that $c_4 = c_5$, hence $c_5 = 1$. Consequently, the expression

$$\sum_{i,j,k} b_i(1 - c_i)a_{ij}a_{jk}c_k \quad (5.5)$$

must be zero (because of $2 \leq k < j < i$). However, by multiplying out and using two fifth-order conditions, the expression in (5.5) should be $1/120$, a contradiction.

The case $p = s = 6$ is treated by considering all “one-leg trees”, i.e., the trees which consist of one leg above the root and the 5th order trees grafted on. The corresponding order conditions have the form

$$\sum_{i,j,\dots} b_i a_{ij} (a_{j,\dots} \dots \text{expressions for order 5}) = \frac{1}{\gamma(t)}.$$

If we let $b'_j = \sum_i b_i a_{ij}$ we are back in the 5th order 5-stage business and can follow the above ideas again. However, the $\gamma(t)$ values are not the same as before; as a consequence, the product UV in (5.3) now becomes

$$UV = \begin{pmatrix} \frac{1!}{(s-2)!} & \frac{2!}{(s-1)!} & 0 \\ \frac{2!}{(s-1)!} & \frac{3!}{s!} & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (s=6). \quad (5.3')$$

Further, for $p = s = 7$ we use the “stork-trees” with order conditions

$$\sum_{i,j,\dots} b_i a_{ij} a_{jk} (a_{k,\dots} \dots \text{expressions for order 5}) = \frac{1}{\gamma(t)}$$

and let $b''_k = \sum_{i,j} b_i a_{ij} a_{jk}$ and so on. The general case $p = s \geq 5$ is now clear. \square

We multiply (5.12) by $1/2$ and then subtract both equations to obtain

$$\sum_{i,j} b_i c_i a_{ij} \left(\sum_k a_{jk} c_k - c_j^2/2 \right) = 0.$$

From (5.7) the parenthesis is zero except when $j = 2$, and therefore

$$\sum_{i=3}^6 b_i c_i a_{i2} = 0 \quad (5.13)$$

replaces (5.11). Our last simplification is to subtract other order conditions from (5.12) to obtain

$$\sum_{i,j} b_i (1 - c_i) a_{ij} c_j (c_j - c_3) = \frac{1}{60} - \frac{c_3}{24}, \quad (5.14)$$

which has fewer terms than before, in particular because $c_6 = 1$ by (5.6) with $j = 6$. The resulting *reduced system* (5.6)-(5.8), (5.10), (5.13), (5.14) can easily be solved as follows:

Algorithm 5.2 (construction of 6-stage 5th order Runge-Kutta methods).

- a) $c_1 = 0$ and $c_6 = 1$ from (5.6) with $j = 6$; c_2, c_3, c_4, c_5 can be chosen as free parameters subject only to some trivial exceptions;
- b) $b_2 = 0$ from (5.8) and b_1, b_3, b_4, b_5, b_6 from the linear system (5.10);
- c) a_{32} from (5.7), $i = 3$; $a_{42} = \lambda$ arbitrary; a_{43} from (5.7), $i = 4$;
- d) a_{52} and a_{62} from the two linear equations (5.13) and (5.6), $j = 2$;
- e) a_{54} from (5.14) and a_{53} from (5.7), $i = 5$;
- f) a_{63}, a_{64}, a_{65} from (5.6), $j = 3, 4, 5$;
- g) finally a_{i1} ($i = 2, \dots, 6$) from (1.9).

Condition (5.6) for $j = 1$ and (5.7) for $i = 6$ are automatically satisfied. This follows as in the proof of Lemma 1.4.

Embedded Formulas of Order 5

Methods of Fehlberg. The methods obtained from Algorithm 5.2 do not all possess an embedded formula of order 4. Fehlberg, interested in the construction of Runge-Kutta pairs of order 4(5), looked mainly for simplifying assumptions which depend only on c_i and a_{ij} , but not on the weights b_i . In this case the simplifying assumptions are useful for the embedded method too. Therefore Fehlberg (1969) considered (5.7), (5.8) and replaced (5.6) by

$$\sum_{j=1}^{i-1} a_{ij} c_j^2 = \frac{c_i^3}{3}, \quad i = 3, \dots, 6. \quad (5.15)$$

As with (5.9) this allows us to disregard all trees of the form $[[\tau, \tau], t_2, \dots, t_m]$. In order that the reduction process of Fig. 5.1 also work on a higher level, we suppose, in addition to $b_2 = 0$, that

$$\sum_i b_i a_{i2} = 0, \quad \sum_i b_i c_i a_{i2} = 0, \quad \sum_{i,j} b_i a_{ij} a_{j2} = 0. \quad (5.16)$$

Then the last equations to be satisfied are

$$\sum_{i,j} b_i a_{ij} c_j^3 = \frac{1}{20} \quad (5.17)$$

and the quadrature conditions (5.10). We remark that the equations (5.7) and (5.15) for $i = 3$ imply

$$c_3 = \frac{3}{2} c_2. \quad (5.18)$$

We now want the method to possess an embedded formula of order 4. Analogously to (5.8) we set $\widehat{b}_2 = 0$. Then conditions (5.7) and (5.15) simplify the conditions of order 4 to 5 linear equations (the 4 quadrature conditions and $\sum_i \widehat{b}_i a_{i2} = 0$) for the 5 unknowns $\widehat{b}_1, \widehat{b}_3, \widehat{b}_4, \widehat{b}_5, \widehat{b}_6$. This system has a second solution (other than the b_i) only if it is singular, which is the case if (see Exercise 1 below)

$$c_4 = \frac{3c_2}{4 - 24c_2 + 45c_2^2}. \quad (5.19)$$

With c_2, c_5, c_6 as free parameters, the above system can be solved and yields an embedded formula of order 4(5). The coefficients of a very popular method, constructed by Fehlberg (1969), are given in Table 5.1.

Table 5.1. Fehlberg 4(5)

0						
$\frac{1}{4}$	$\frac{1}{4}$					
$\frac{3}{8}$	$\frac{3}{32}$	$\frac{9}{32}$				
$\frac{12}{13}$	$\frac{1932}{2197}$	$-\frac{7200}{2197}$	$\frac{7296}{2197}$			
1	$\frac{439}{216}$	-8	$\frac{3680}{513}$	$-\frac{845}{4104}$		
$\frac{1}{2}$	$-\frac{8}{27}$	2	$-\frac{3544}{2565}$	$\frac{1859}{4104}$	$-\frac{11}{40}$	
y_1	$\frac{25}{216}$	0	$\frac{1408}{2565}$	$\frac{2197}{4104}$	$-\frac{1}{5}$	0
\widehat{y}_1	$\frac{16}{135}$	0	$\frac{6656}{12825}$	$\frac{28561}{56430}$	$-\frac{9}{50}$	$\frac{2}{55}$

All of the methods of Fehlberg are of the type $p(\hat{p})$ with $p < \hat{p}$. Hence, the lower order approximation is intended to be used as initial value for the next step. In order to make his methods optimal, Fehlberg tried to minimize the error coefficients for the lower order result y_1 . This has the disadvantage that the local extrapolation mode (continue the integration with the higher order result) does not make sense and the estimated “error” can become substantially smaller than the true error.

It is possible to do a lot better than the pair of Fehlberg currently
regarded as “best.” (L.F. Shampine 1986)

Table 5.2. Dormand-Prince 5(4) (DOPRI5)

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
y_1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
\hat{y}_1	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

Dormand & Prince pairs. The first efforts at minimizing the error coefficients of the higher order result, which is then used as numerical solution, were undertaken by Dormand & Prince (1980). Their methods of order 5 are constructed with the help of Algorithm 5.2 under the additional hypothesis (5.15). This condition is achieved by fixing the parameters c_3 and a_{42} in such a way that (5.15) holds for $i = 3$ and $i = 4$. The remaining two relations ($i = 5, 6$) are then automatically satisfied. To see this, multiply the difference $e_i = \sum_{j=1}^{i-1} a_{ij}c_j^2 - c_i^3/3$ by b_i and $b_i c_i$, respectively, sum up and deduce that all e_i must vanish.

In order to equip the method with an embedded formula, Dormand & Prince propose to use the FSAL idea (i.e., add y_1 as 7th stage). In this way the restriction (5.19) for c_4 is no longer necessary. We fix arbitrarily $\hat{b}_7 \neq 0$, put $\hat{b}_2 = 0$ (as in (5.8)) and compute the remaining \hat{b}_i , as above for the Fehlberg case from the 4 quadrature conditions and from $\sum_i \hat{b}_i a_{i2} = 0$.

We have thus obtained a family of 5th order Runge-Kutta methods with 4th

order embedded solution with c_2, c_4, c_5 as free parameters. Dormand & Prince (1980) have undertaken an extensive search to determine these parameters in order to minimize the error coefficients for y_1 and found that $c_2 = 1/5$, $c_4 = 4/5$ and $c_5 = 8/9$ was a close rational approximation to an optimal choice. Table 5.2 presents the coefficients of this method. The corresponding code of the Appendix is called DOPRI5.

Higher Order Processes

Order 6. By Theorem 5.1 at least 7 stages are necessary for order 6. A. Huřa (1956) constructed 6th order processes with 8 stages. Finally, methods with $s=7$, the optimal number, were derived by Butcher (1964b) along similar lines as above. He arrived at an algorithm where c_2, c_3, c_5, c_6 are free parameters.

Order 7. The existence of such a method with 8 stages is impossible by the following barrier:

Theorem 5.3 (Butcher 1965b). *For $p \geq 7$ no explicit Runge-Kutta method exists of order p with $s = p + 1$ stages.*

Since the proof of this theorem is much more complicated than that of Theorem 5.1, we do not reproduce it here.

This raises the question, whether 7th order methods with 9 stages exist. Such methods, announced by Butcher (1965b), do exist; see Verner (1971).

Order 8. As to methods of order 8, Curtis (1970) and Cooper & Verner (1972) have constructed such processes with $s = 11$. It was for a long time an open question whether there exist methods with 10 stages. John Butcher's dream of settling this difficult question before his 50th birthday did not become true. But he finally succeeded in proving the non-existence for Dahlquist's 60th birthday:

Theorem 5.4 (Butcher 1985b). *For $p \geq 8$ no explicit Runge-Kutta method exists of order p with $s = p + 2$ stages.*

For the proof, which is still more complicated, we again refer to Butcher's original paper.

Order 10. These are the highest order explicitly constructed explicit Runge-Kutta methods. Curtis (1975) constructed an 18-stage method of order 10. His construction was based solely on simplifying assumptions of the type (5.7), (5.8) and their extensions. Hairer (1978) then constructed a 17-stage method by using the complete arsenal of simplifying ideas. For more details, see the first edition, p. 189.

Embedded Formulas of High Order

It was mainly the formula manipulation genius Fehlberg who first derived high order embedded formulas. His greatest success was his 7th order formula with 8th order error estimate (Fehlberg 1968) which is of frequent use in all high precision computations, e.g., in astronomy. The coefficients are reproduced in Table 5.3.

Table 5.3. Fehlberg 7(8)													
0													
$\frac{2}{27}$	$\frac{2}{27}$												
$\frac{1}{9}$	$\frac{1}{36}$	$\frac{1}{12}$											
$\frac{1}{6}$	$\frac{1}{24}$	0	$\frac{1}{8}$										
$\frac{5}{12}$	$\frac{5}{12}$	0	$-\frac{25}{16}$	$\frac{25}{16}$									
$\frac{1}{2}$	$\frac{1}{20}$	0	0	$\frac{1}{4}$	$\frac{1}{5}$								
$\frac{5}{6}$	$-\frac{25}{108}$	0	0	$\frac{125}{108}$	$-\frac{65}{27}$	$\frac{125}{54}$							
$\frac{1}{6}$	$\frac{31}{300}$	0	0	0	$\frac{61}{225}$	$-\frac{2}{9}$	$\frac{13}{900}$						
$\frac{2}{3}$	2	0	0	$-\frac{53}{6}$	$\frac{704}{45}$	$-\frac{107}{9}$	$\frac{67}{90}$	3					
$\frac{1}{3}$	$-\frac{91}{108}$	0	0	$\frac{23}{108}$	$-\frac{976}{135}$	$\frac{311}{54}$	$-\frac{19}{60}$	$\frac{17}{6}$	$-\frac{1}{12}$				
1	$\frac{2383}{4100}$	0	0	$-\frac{341}{164}$	$\frac{4496}{1025}$	$-\frac{301}{82}$	$\frac{2133}{4100}$	$\frac{45}{82}$	$\frac{45}{164}$	$\frac{18}{41}$			
0	$\frac{3}{205}$	0	0	0	0	$-\frac{6}{41}$	$-\frac{3}{205}$	$-\frac{3}{41}$	$\frac{3}{41}$	$\frac{6}{41}$	0		
1	$-\frac{1777}{4100}$	0	0	$-\frac{341}{164}$	$\frac{4496}{1025}$	$-\frac{289}{82}$	$\frac{2193}{4100}$	$\frac{51}{82}$	$\frac{33}{164}$	$\frac{19}{41}$	0	1	
y_1	$\frac{41}{840}$	0	0	0	0	$\frac{34}{105}$	$\frac{9}{35}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{9}{280}$	$\frac{41}{840}$	0	0
\hat{y}_1	0	0	0	0	0	$\frac{34}{105}$	$\frac{9}{35}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{9}{280}$	0	$\frac{41}{840}$	$\frac{41}{840}$

Fehlberg's methods suffer from the fact that they give identically zero error estimates for quadrature problems $y' = f(x)$. The first high order embedded formulas which avoid this drawback were constructed by Verner (1978). One of Verner's methods (see Table 5.4) has been implemented by T.E. Hull, W.H. Enright and K.R. Jackson as DVERK and is widely used.

Table 5.4. Verner's method of order 6(5) (DVERK)

0								
$\frac{1}{6}$	$\frac{1}{6}$							
$\frac{4}{15}$	$\frac{4}{15}$	$\frac{16}{75}$						
$\frac{2}{3}$	$\frac{5}{6}$	$-\frac{8}{3}$	$\frac{5}{2}$					
$\frac{5}{6}$	$-\frac{165}{64}$	$\frac{55}{6}$	$-\frac{425}{64}$	$\frac{85}{96}$				
1	$\frac{12}{5}$	-8	$\frac{4015}{612}$	$-\frac{11}{36}$	$\frac{88}{255}$			
$\frac{1}{15}$	$-\frac{8263}{15000}$	$\frac{124}{75}$	$-\frac{643}{680}$	$-\frac{81}{250}$	$\frac{2484}{10625}$	0		
1	$\frac{3501}{1720}$	$-\frac{300}{43}$	$\frac{297275}{52632}$	$-\frac{319}{2322}$	$\frac{24068}{84065}$	0	$\frac{3850}{26703}$	
y_1	$\frac{3}{40}$	0	$\frac{875}{2244}$	$\frac{23}{72}$	$\frac{264}{1955}$	0	$\frac{125}{11592}$	$\frac{43}{616}$
\hat{y}_1	$\frac{13}{160}$	0	$\frac{2375}{5984}$	$\frac{5}{16}$	$\frac{12}{85}$	$\frac{3}{44}$	0	0

An 8th Order Embedded Method

The first high order methods with small error constants of the *higher* order solution were constructed by Prince & Dormand (1981, Code DOPRI8 of the first edition). In the following we describe the construction of a new Dormand & Prince pair of order 8(6) which will also allow a cheap and accurate dense output (see Section II.6). This method has been announced, but not published, in Dormand & Prince (1989, p. 983). We are grateful to P. Prince for mailing us the coefficients and for his help in recovering their construction.

The essential difficulty for the construction of a high order Runge-Kutta method is to set up a “good” reduced system which implies all order conditions of Theorem 2.13. At the same time it should be simple enough to be easily solved. In extending the ideas for the construction of a 5th order process (see above), Dormand & Prince proceed as follows:

Reduced system. Suppose $s = 12$ and consider for the coefficients c_i , b_i and a_{ij} the equations:

$$\sum_{i=1}^s b_i c_i^{q-1} = 1/q, \quad q = 1, \dots, 8 \quad (5.20a)$$

$$\sum_{j=1}^{i-1} a_{ij} = c_i, \quad i = 1, \dots, s \quad (5.20b)$$

$$\sum_{j=1}^{i-1} a_{ij} c_j = c_i^2/2, \quad i = 3, \dots, s \quad (5.20c)$$

$$\sum_{j=1}^{i-1} a_{ij} c_j^2 = c_i^3/3, \quad i = 3, \dots, s \quad (5.20d)$$

$$\sum_{j=1}^{i-1} a_{ij} c_j^3 = c_i^4/4, \quad i = 6, \dots, s \quad (5.20e)$$

$$\sum_{j=1}^{i-1} a_{ij} c_j^4 = c_i^5/5, \quad i = 6, \dots, s \quad (5.20f)$$

$$b_2 = b_3 = b_4 = b_5 = 0 \quad (5.20g)$$

$$a_{i2} = 0 \quad \text{for } i \geq 4, \quad a_{i3} = 0 \quad \text{for } i \geq 6 \quad (5.20h)$$

$$\sum_{i=j+1}^s b_i a_{ij} = b_j(1 - c_j), \quad j = 4, 5, 10, 11, 12 \quad (5.20i)$$

$$\sum_{i=j+1}^s b_i c_i a_{ij} = 0, \quad j = 4, 5 \quad (5.20j)$$

$$\sum_{i=j+1}^s b_i c_i^2 a_{ij} = 0, \quad j = 4, 5 \quad (5.20k)$$

$$\sum_{i=k+2}^s b_i c_i \sum_{j=k+1}^{i-1} a_{ij} a_{jk} = 0, \quad k = 4, 5 \quad (5.20l)$$

$$\sum_{i=1}^s b_i c_i \sum_{j=1}^{i-1} a_{ij} c_j^5 = 1/48. \quad (5.20m)$$

Verification of the order conditions. The equations (5.20a) are the order conditions for the bushy trees $[\tau, \dots, \tau]$ and (5.20m) is that for the tree $[\tau, [\tau, \tau, \tau, \tau, \tau]]$. For the verification of further order conditions we shall show that the reduced system implies

$$\sum_{i=j+1}^s b_i a_{ij} = b_j(1 - c_j) \quad \text{for all } j. \quad (5.21)$$

If we denote the difference by $d_j = \sum_{i=j+1}^s b_i a_{ij} - b_j(1 - c_j)$ then $d_2 = d_3 = 0$ by (5.20g,h) and $d_4 = d_5 = d_{10} = d_{11} = d_{12} = 0$ by (5.20i). The conditions (5.20a-g) imply

$$\sum_{j=1}^s d_j c_j^{q-1} = 0 \quad \text{for } q = 1, \dots, 5. \quad (5.22)$$

Hence, the remaining 5 values must also vanish if c_1, c_6, c_7, c_8, c_9 are distinct. The significance of condition (5.21) is already known from Lemma 1.3 and from formula (5.6). It implies that all one-leg trees $t = [t_1]$ can be disregarded.



Fig. 5.2. Use of simplifying assumptions

Conditions (5.20c-f) are an extension of (5.6) and (5.15). Their importance will be, once more, demonstrated on an example. Consider the two trees of Fig. 5.2 and suppose that their encircled parts are identical. Then the corresponding order

conditions are

$$\sum_{i,j=1}^s \Theta_i a_{ij} c_j^3 = \frac{1}{r \cdot 5 \cdot 4} \quad \text{and} \quad \sum_{i=1}^s \Theta_i c_i^4 = \frac{1}{r \cdot 5} \quad (5.23)$$

with known values for Θ_i and r . If (5.20e) is satisfied and if

$$\Theta_2 = \Theta_3 = \Theta_4 = \Theta_5 = 0 \quad (5.24)$$

then both conditions are equivalent so that the left-hand tree can be neglected. The conditions (5.20g,i-1) correspond to (5.24) for certain trees. Finally the assumption (5.20h) together with (5.20g,i-k) implies that for arbitrary Φ_i , Ψ_j and for $q \in \{1, 2, 3\}$,

$$\begin{aligned} \sum_i b_i \Phi_i a_{i2} &= 0 \\ \sum_{i,j} b_i \Phi_i a_{ij} \Psi_j a_{j2} &= 0 \\ \sum_{i,j,k} b_i c_i^{q-1} a_{ij} \Phi_j a_{jk} \Psi_k a_{k2} &= 0 \end{aligned} \quad \text{and} \quad \begin{aligned} \sum_i b_i \Phi_i a_{i3} &= 0 \\ \sum_{i,j} b_i c_i^{q-1} a_{ij} \Phi_j a_{j3} &= 0 \end{aligned}$$

which are again conditions of type (5.24). Using these relations the verification of the order conditions (order 8) is straightforward; all trees are reduced to those corresponding to (5.20a) and (5.20m).

Solving the reduced system. Compared to the original 200 order conditions of Theorem 2.13 for the 78 coefficients b_i, a_{ij} (the c_i are defined by (5.20b)), the 74 conditions of the reduced system present a considerable simplification. We can hope for a solution with 4 degrees of freedom.

We start by expressing the coefficients b_i, a_{ij} in terms of the c_i . Because of (5.20g), condition (5.20a) represents a linear system for b_1, b_6, \dots, b_{12} , which has a unique solution if c_1, c_6, \dots, c_{12} are distinct. For a fixed i ($1 \leq i \leq 8$) conditions (5.20b-f) represent a linear system for $a_{i1}, \dots, a_{i,i-1}$. Since there are sometimes less unknowns than equations (mainly due to (5.20h)) restrictions have to be imposed on the c_i . One verifies (similarly to (5.18)) that the relations

$$\begin{aligned} c_1 &= 0, & c_2 &= \frac{2}{3} c_3, & c_3 &= \frac{2}{3} c_4, \\ c_4 &= \frac{6 - \sqrt{6}}{10} c_6, & c_5 &= \frac{6 + \sqrt{6}}{10} c_6, & c_6 &= \frac{4}{3} c_7 \end{aligned} \quad (5.25a)$$

allow the computation of the a_{ij} with $i \leq 8$ (Step 1 in Fig. 5.3).

If $b_{12} \neq 0$ (which will be assumed in our construction), condition (5.20i) for $j = 12$ implies

$$c_{12} = 1, \quad (5.25b)$$

and for $j = 11$ it yields the value for $a_{12,11}$. We next compute the expressions

$$e_j = \sum_{i=j+1}^s b_i c_i a_{ij} - \frac{b_j}{2} (1 - c_j^2), \quad j = 1, \dots, s. \quad (5.26)$$

& Prince propose the following numerical values:

$$c_7 = 1/4, \quad c_8 = 4/13, \quad c_{10} = 3/5, \quad c_{11} = 6/7.$$

All remaining coefficients are then determined by the above procedure. Since c_4 and c_5 (see (5.25a)) are not rational, there is no easy way to present the coefficients in a tableau.

Embedded method. We look for a second method with the same c_i, a_{ij} but with different weights, say \hat{b}_i . If we require that

$$\sum_{i=1}^s \hat{b}_i c_i^{q-1} = 1/q, \quad q = 1, \dots, 6 \quad (5.29a)$$

$$\hat{b}_2 = \hat{b}_3 = \hat{b}_4 = \hat{b}_5 = 0 \quad (5.29b)$$

$$\sum_{i=j+1}^s \hat{b}_i a_{ij} = 0, \quad j = 4, 5 \quad (5.29c)$$

then one can verify (similarly as above for the 8th order method) that the corresponding Runge-Kutta method is of order 6. The system (5.29) consists of 12 linear equations for 12 unknowns. A comparison with (5.20) shows that b_1, \dots, b_{12} is a solution of (5.29). Furthermore, the corresponding homogeneous system has the nontrivial solution e_1, \dots, e_{12} (see (5.27) and (5.20)). Therefore

$$\hat{b}_i = b_i + \alpha e_i \quad (5.30)$$

is a solution of (5.29) for all values of α . Dormand & Prince suggest taking α in such a way that $\hat{b}_6 = 2$.

A program based on this method (with a different error estimator, see Section II.10) has been written and is called DOP853. It is documented in the Appendix. The performance of this code, compared to methods of lower order, is impressive. See for example the results for the Brusselator in Fig. 4.2.

Exercises

1. Consider a Runge-Kutta method with s stages that satisfies (5.7)-(5.8), (5.15), (5.17) and the first two relations of (5.16).
 - a) If the relation (5.19) holds, then the method possesses an embedded formula of order 4.
 - b) The condition (5.19) implies that the last relation of (5.16) is automatically satisfied.

Hint. The order conditions for the embedded method constitute a linear system for the \hat{b}_i which has to be singular. This implies that

$$a_{i2} = \alpha c_i + \beta c_i^2 + \gamma c_i^3 \quad \text{for} \quad i \neq 2. \quad (5.31)$$

Multiplying (5.31) with b_i and $b_i c_i$ and summing up, yields two relations for $\alpha, \beta, J\gamma$. These together with (5.31) for $i = 3, 4$ yield (5.19).

- Construct a 6-stage 5th order formula with $c_3 = 1/3$, $c_4 = 1/2$, $c_5 = 2/3$ possessing an embedded formula of order 4.
- (Butcher). Show that for any Runge-Kutta method of order 5,

$$\sum_i b_i \left(\sum_j a_{ij} c_j - \frac{c_i^2}{2} \right)^2 = 0.$$

Consequently, there exists no explicit Runge-Kutta method of order 5 with all $b_i > 0$.

Hint. Multiply out and use order conditions.

- Write a code with a high order Runge-Kutta method (or take one) and solve numerically the Arenstorf orbit of the restricted three body problem (0.1) (see the introduction) with initial values

$$\begin{aligned} y_1(0) &= 0.994, & y_1'(0) &= 0, & y_2(0) &= 0, \\ y_2'(0) &= -2.0317326295573368357302057924, \end{aligned}$$

Compute the solutions for

$$x_{\text{end}} = 11.124340337266085134999734047.$$

The initial values are chosen such that the solution is periodic to this precision. The plotted solution curve has one loop less than that of the introduction.

- (Shampine 1979). Show that the storage requirement of a Runge-Kutta method can be substantially decreased if s is large.

Hint. Suppose, for example, that $s = 15$.

After computing (see (1.8)) k_1, k_2, \dots, k_9 , compute the sums

$$\sum_{j=1}^9 a_{ij} k_j \quad \text{for } i = 10, 11, 12, 13, 14, 15, \quad \sum_{j=1}^9 b_j k_j, \quad \sum_{j=1}^9 \hat{b}_j k_j;$$

then the memories occupied by k_2, k_3, \dots, k_9 are not needed any longer. Another possibility for reducing the memory requirement is offered by the zero-pattern of the coefficients.

- Show that the reduced system (5.20) implies (5.25c).

Hint. The equations (5.20b-f) imply that for $i \in \{1, 6, 7, 8, 9\}$

$$\alpha a_{i4} + \beta a_{i5} = \sigma_3 \frac{c_i^2}{2} - \sigma_2 \frac{c_i^3}{3} + \sigma_1 \frac{c_i^4}{4} - \frac{c_i^5}{5} \quad (5.32)$$

with σ_j given by (5.28). The constants α and β are not important. Further, for the same values of i one has

$$\begin{aligned} 0 &= c_i(c_i - c_6)(c_i - c_7)(c_i - c_8)(c_i - c_9) \\ &= \sigma_3 c_9 c_i - (\sigma_3 + c_9 \sigma_2) c_i^2 + (\sigma_2 + c_9 \sigma_1) c_i^3 - (\sigma_1 + c_9) c_i^4 + c_i^5. \end{aligned} \quad (5.33)$$

Multiplying (5.32) and (5.33) by $e_i, b_i, b_i c_i, b_i c_i^2$, summing up from $i = 1$ to s and using (5.20) gives the relation

$$\begin{pmatrix} \times & \times & \times \\ \times & \times & \times \\ 0 & 0 & b_{12}^{-1} \end{pmatrix} \begin{pmatrix} e_{10} & b_{10} & b_{10} c_{10} & b_{10} c_{10}^2 \\ e_{11} & b_{11} & b_{11} c_{11} & b_{11} c_{11}^2 \\ 0 & b_{12} & b_{12} & b_{12} \end{pmatrix} = \begin{pmatrix} 0 & \gamma_1 & \gamma_2 & \gamma_3 \\ 0 & \delta_1 & \delta_2 & \delta_3 \\ 0 & 1 & 1 & 1 \end{pmatrix} \quad (5.34)$$

where

$$\begin{aligned} \gamma_j &= \frac{\sigma_3}{2 \cdot (j+2)} - \frac{\sigma_2}{3 \cdot (j+3)} + \frac{\sigma_1}{4 \cdot (j+4)} - \frac{1}{5 \cdot (j+5)} \\ \delta_j &= \frac{\sigma_3 c_9}{j+1} - \frac{\sigma_3 + c_9 \sigma_2}{j+2} + \frac{\sigma_2 + c_9 \sigma_1}{j+3} - \frac{\sigma_1 + c_9}{j+4} + \frac{1}{j+5} \end{aligned}$$

and the “ \times ” indicate certain values. Deduce from (5.34) and $e_{11} \neq 0$ that the most left matrix of (5.34) is singular. This implies that the right-hand matrix of (5.34) is of rank 2 and yields equation (5.25c).

7. Prove that the 8th order method given by (5.20; $s = 12$) does not possess a 6th order embedding with $\hat{b}_{12} \neq b_{12}$, not even if one adds the numerical result y_1 as 13th stage (FSAL).

II.6 Dense Output, Discontinuities, Derivatives

... providing “interpolation” for Runge-Kutta methods. ... this capability and the features it makes possible will be the hallmark of the next generation of Runge-Kutta codes.

(L.F. Shampine 1986)

The present section is mainly devoted to the construction of dense output formulas for Runge-Kutta methods. This is important for many practical questions such as graphical output, event location or the treatment of discontinuities in differential equations. Further, the numerical computation of derivatives with respect to initial values and parameters is discussed, which is particularly useful for the integration of boundary value problems.

Dense Output

Classical Runge-Kutta methods are inefficient, if the number of output points becomes very large (Shampine, Watts & Davenport 1976). This motivated the construction of dense output formulas (Horn 1983). These are Runge-Kutta methods which provide, in addition to the numerical result y_1 , cheap numerical approximations to $y(x_0 + \theta h)$ for the *whole* integration interval $0 \leq \theta \leq 1$. “Cheap” means without or, at most, with only a few additional function evaluations.

We start from an s -stage Runge-Kutta method with given coefficients c_i, a_{ij} and b_j , eventually add $s^* - s$ new stages, and consider formulas of the form

$$u(\theta) = y_0 + h \sum_{i=1}^{s^*} b_i(\theta) k_i, \quad (6.1)$$

where

$$k_i = f\left(x_0 + c_i h, y_0 + h \sum_{j=1}^{i-1} a_{ij} k_j\right), \quad i = 1, \dots, s^* \quad (6.2)$$

and $b_i(\theta)$ are polynomials to be determined such that

$$u(\theta) - y(x_0 + \theta h) = \mathcal{O}(h^{p^*+1}). \quad (6.3)$$

Usually $s^* \geq s + 1$ since we include (at least) the first function evaluation of the subsequent step $k_{s+1} = hf(x_0 + h, y_1)$ in the formula with $a_{s+1,j} = b_j$ for all j . A Runge-Kutta method, provided with a formula (6.1), will be called a *continuous* Runge-Kutta method.

Theorem 6.1. *The error of the approximation (6.1) is of order p^* (i.e., the local error satisfies (6.3)), if and only if*

$$\sum_{j=1}^{s^*} b_j(\theta) \Phi_j(t) = \frac{\theta^{\varrho(t)}}{\gamma(t)} \quad \text{for} \quad \varrho(t) \leq p^* \quad (6.4)$$

with $\Phi_j(t)$, $\varrho(t)$, $\gamma(t)$ given in Section II.2.

Proof. The q th derivative (with respect to h) of the numerical approximation is given by (2.14) with b_j replaced by $b_j(\theta)$; that of the exact solution $y(x_0 + \theta h)$ is $\theta^q y^{(q)}(x_0)$. The statement thus follows as in Theorem 2.13. \square

Corollary 6.2. *Condition (6.4) implies that the derivatives of (6.1) approximate the derivatives of the exact solution as*

$$h^{-k} u^{(k)}(\theta) - y^{(k)}(x_0 + \theta h) = \mathcal{O}(h^{p^* - k + 1}). \quad (6.5)$$

Proof. Comparing the q th derivative (with respect to h) of $u'(\theta)$ with that of $hy'(x_0 + \theta h)$ we find that (6.5) (for $k = 1$) is equivalent to

$$\sum_{j=1}^{s^*} b'_j(\theta) \Phi_j(t) = \frac{\varrho(t) \theta^{\varrho(t)-1}}{\gamma(t)} \quad \text{for} \quad \varrho(t) \leq p^*.$$

This, however, follows from (6.4) by differentiation. The case $k > 1$ is obtained similarly. \square

We write the polynomials $b_j(\theta)$ as

$$b_j(\theta) = \sum_{q=1}^{p^*} b_{jq} \theta^q, \quad (6.6)$$

so that the equations (6.4) become a system of simultaneous linear equations of the form

$$\underbrace{\begin{pmatrix} 1 & 1 & \dots & 1 \\ \Phi_1(t_{21}) & \Phi_2(t_{21}) & \dots & \Phi_{s^*}(t_{21}) \\ \Phi_1(t_{31}) & \Phi_2(t_{31}) & \dots & \Phi_{s^*}(t_{31}) \\ \vdots & \vdots & & \vdots \end{pmatrix}}_{\Phi} \underbrace{\begin{pmatrix} b_{11} & b_{12} & b_{13} & \dots \\ b_{21} & b_{22} & b_{23} & \dots \\ \vdots & \vdots & \vdots & \vdots \\ b_{s^*1} & b_{s^*2} & b_{s^*3} & \dots \end{pmatrix}}_B = \underbrace{\begin{pmatrix} 1 & 0 & 0 & \dots \\ 0 & \frac{1}{2} & 0 & \dots \\ 0 & 0 & \frac{1}{3} & \dots \\ 0 & 0 & \frac{1}{6} & \dots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}}_G \quad (6.4')$$

where the $\Phi_j(t)$ are known numbers depending on a_{ij} and c_i . Using standard linear algebra the solution of this system can easily be discussed. It may happen,

however, that the order p^* of the dense output is smaller than the order p of the underlying method.

Example. For “the” Runge-Kutta method of Table 1.2 (with $s^* = s = 4$) equations (6.4') with $p^* = 3$ produce a unique solution

$$b_1(\theta) = \theta - \frac{3\theta^2}{2} + \frac{2\theta^3}{3}, \quad b_2(\theta) = b_3(\theta) = \theta^2 - \frac{2\theta^3}{3}, \quad b_4(\theta) = -\frac{\theta^2}{2} + \frac{2\theta^3}{3}$$

which constitutes a dense output solution which is globally continuous but not C^1 .

Hermite interpolation. A much easier way (than solving (6.4')) and more efficient for low order dense output formulas is the use of Hermite interpolation (Shampine 1985). Whatever the method is, we have two function values y_0, y_1 and two derivatives $f_0 = f(x_0, y_0)$, $f_1 = f(x_0 + h, y_1)$ at our disposal and can thus do cubic polynomial interpolation. The resulting formula is

$$u(\theta) = (1 - \theta)y_0 + \theta y_1 + \theta(\theta - 1) \left((1 - 2\theta)(y_1 - y_0) + (\theta - 1)hf_0 + \theta hf_1 \right). \quad (6.7)$$

Inserting the definition of y_1 into (6.7) shows that Hermite interpolation is a special case of (6.1). Whenever the underlying method is of order $p \geq 3$ we thus obtain a continuous Runge-Kutta method of order 3.

Since the function and derivative values on the right side of the first interval coincide with those on the left side of the second interval, Hermite interpolation leads to a globally C^1 approximation of the solution.

The 4-stage 4th order methods of Section II.1 do not possess a dense output of order 4 without any additional function evaluations (see Exercise 1). Therefore the question arises whether it is really important to have a dense output of the same order. Let us consider an interval far away from the initial value, say $[x_n, x_{n+1}]$, and denote by $z(x)$ the local solution, i.e., the solution of the differential equation which passes through (x_n, y_n) . Then the error of the dense output is composed of two terms:

$$u(\theta) - y(x_n + \theta h) = (u(\theta) - z(x_n + \theta h)) + (z(x_n + \theta h) - y(x_n + \theta h)).$$

The term to the far right reflects the global error of the method and is of size $\mathcal{O}(h^p)$. In order that both terms be of the same order of magnitude it is thus sufficient to require $p^* = p - 1$.

The situation changes, if we also need accurate values of the derivative $y'(x_n + \theta h)$ (see Section 5 of Enright, Jackson, Nørsett & Thomsen (1986) for a discussion of problems where this is important). We have

$$h^{-1}u'(\theta) - y'(x_n + \theta h) = (h^{-1}u'(\theta) - z'(x_n + \theta h)) + (z'(x_n + \theta h) - y'(x_n + \theta h))$$

and the term to the far right is of size $\mathcal{O}(h^p)$ if $f(x, y)$ satisfies a Lipschitz condition. A comparison with (6.5) shows that we need $p^* = p$ in order that both error terms be of comparable size.

Boot-strapping process (Enright, Jackson, Nørsett & Thomsen 1986). This is a general procedure for increasing iteratively the order of dense output formulas.

Suppose that we already have a 3rd order dense output at our disposal (e.g., from Hermite interpolation). We then fix arbitrarily an $\alpha \in (0, 1)$ and denote the 3rd order approximation at $x_0 + \alpha h$ by y_α . The idea is now that $hf(x_0 + \alpha h, y_\alpha)$ is a 4th order approximation to $hy'(x_0 + \alpha h)$. Consequently, the 4th degree polynomial $u(\theta)$ defined by

$$\begin{aligned} u(0) &= y_0, & u'(0) &= hf(x_0, y_0) \\ u(1) &= y_1, & u'(1) &= hf(x_0 + h, y_1) \\ & & u'(\alpha) &= hf(x_0 + \alpha h, y_\alpha) \end{aligned} \quad (6.8)$$

(which exists uniquely for $\alpha \neq 1/2$) yields the desired formula. The interpolation error is $\mathcal{O}(h^5)$ and each quantity of (6.8) approximates the corresponding exact solution value with an error of $\mathcal{O}(h^5)$.

The extension to arbitrary order is straightforward. Suppose that a dense output formula $u_0(\theta)$ of order $p^* < p$ is known. We then evaluate this polynomial at $p^* - 2$ distinct points $\alpha_i \in (0, 1)$ and compute the values $f(x_0 + \alpha_i h, u_0(\alpha_i))$. The interpolation polynomial $u_1(\theta)$ of degree $p^* + 1$, defined by

$$\begin{aligned} u_1(0) &= y_0, & u_1'(0) &= hf(x_0, y_0) \\ u_1(1) &= y_1, & u_1'(1) &= hf(x_0 + h, y_1) \\ u_1'(\alpha_i) &= hf(x_0 + \alpha_i h, u_0(\alpha_i)), & i &= 1, \dots, p^* - 2, \end{aligned} \quad (6.9)$$

yields an interpolation formula of order $p^* + 1$. Obviously, the α_i in (6.9) have to be chosen such that the corresponding interpolation problem admits a solution.

Continuous Dormand & Prince Pairs

The method of Dormand & Prince (Table 5.2) is of order 5(4) so that we are mainly interested in dense output formulas with $p^* = 4$ and $p = 5$.

Order 4. A continuous formula of order 4 can be obtained without any additional function evaluation. Since the coefficients satisfy (5.7), it follows from the difference of the order conditions for the trees t_{31} and t_{32} (notation of Table 2.2) that

$$b_2(\theta) = 0 \quad (6.10)$$

is necessary. This condition together with (5.7) and (5.15) then implies that the order conditions are equivalent for the following pairs of trees: t_{31} and t_{32} , t_{41} and t_{42} , t_{41} and t_{43} . Hence, for order 4, only 5 conditions have to be considered (the four quadrature conditions and $\sum_i b_i(\theta)a_{i2} = 0$). We can arbitrarily choose $b_7(\theta)$ and the coefficients $b_1(\theta), b_3(\theta), \dots, b_6(\theta)$ are then uniquely determined.

As for the choice of $b_7(\theta)$, Shampine (1986) proposed minimizing, for each θ , the error coefficients (Theorem 3.2)

$$e(t) = \theta^5 - \gamma(t) \sum_{j=1}^7 b_j(\theta) \Phi_j(t) \quad \text{for } t \in T_5, \quad (6.11)$$

weighted by $\alpha(t)$ of Definition 2.5, in the square norm. These expressions can be seen to depend linearly on $b_7(\theta)$,

$$\alpha(t)e(t) = \zeta(t, \theta) - b_7(\theta)\eta(t),$$

thus the minimal value is found for

$$b_7(\theta) = \sum_{t \in T_5} \zeta(t, \theta) \eta(t) / \sum_{t \in T_5} \eta^2(t).$$

The resulting formula, given by Dormand & Prince (1986), is

$$b_7(\theta) = \theta^2(\theta - 1) + \theta^2(\theta - 1)^2 10 \cdot (7414447 - 829305\theta) / 29380423. \quad (6.12)$$

The other coefficients, written in a fashion which makes the Hermite-part clearly visible, are then given by

$$\begin{aligned} b_1(\theta) &= \theta^2(3 - 2\theta) \cdot b_1 + \theta(\theta - 1)^2 \\ &\quad - \theta^2(\theta - 1)^2 5 \cdot (2558722523 - 31403016\theta) / 11282082432 \\ b_3(\theta) &= \theta^2(3 - 2\theta) \cdot b_3 + \theta^2(\theta - 1)^2 100 \cdot (882725551 - 15701508\theta) / 32700410799 \\ b_4(\theta) &= \theta^2(3 - 2\theta) \cdot b_4 - \theta^2(\theta - 1)^2 25 \cdot (443332067 - 31403016\theta) / 1880347072 \\ b_5(\theta) &= \theta^2(3 - 2\theta) \cdot b_5 + \theta^2(\theta - 1)^2 32805 \cdot (23143187 - 3489224\theta) / 199316789632 \\ b_6(\theta) &= \theta^2(3 - 2\theta) \cdot b_6 - \theta^2(\theta - 1)^2 55 \cdot (29972135 - 7076736\theta) / 822651844. \end{aligned} \quad (6.13)$$

It can be directly verified that the interpolation polynomial $u(\theta)$ defined by (6.10), (6.12) and (6.13) satisfies

$$\begin{aligned} u(0) &= y_0, & u'(0) &= hf(x_0, y_0), \\ u(1) &= y_1, & u'(1) &= hf(x_0 + h, y_1), \end{aligned} \quad (6.14)$$

so that it produces globally a C^1 approximation of the solution.

Instead of using the above 5th degree polynomial $u(\theta)$, Shampine (1986) suggests evaluating it only at the midpoint, $y_{1/2} = u(1/2)$, and then doing quartic polynomial interpolation with the five values y_0 , $hf(x_0, y_0)$, y_1 , $hf(x_0 + h, y_1)$, $y_{1/2}$. This dense output is also C^1 , is easier to implement and the difference to the above formula "... is not significant" (Dormand & Prince 1986).

We have implemented Shampine's dense output in the code DOPRI5 (see Appendix). The advantages of such a dense output for graphical representations of the solution can already be seen from Fig. 0.1 of the introduction to Chapter II. For a more thorough study we have applied DOPRI5 to the Brusselator (4.15) with initial

values $y_1(0) = 1.5$, $y_2(0) = 3$, integration interval $0 \leq x \leq 10$ and error tolerance $Atol = Rtol = 10^{-4}$. The global error of the above 4th order continuous solution is displayed in Fig. 6.1 for both components. The error shows the same quality throughout; the grid points, which are represented by the symbols \square and \circ , are by no means outstanding.

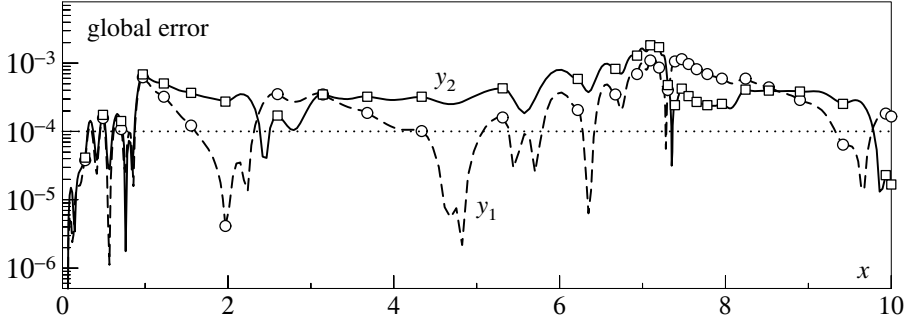


Fig. 6.1. Error of dense output of DOPRI5

Order 5. For a dense output of order $p^* = 5$ for the Dormand & Prince method the linear system (6.4') has no solution since

$$\text{rank}(\Phi|G) = 9 \quad \text{and} \quad \text{rank}(\Phi) = 7 \quad (6.15)$$

as can be verified by Gaussian elimination. Such a linear system has a solution if and only if the two ranks in (6.15) are *equal*. So we must append additional stages to the method. Each new stage adds a new column to the matrix Φ , thus may increase the rank of Φ by one without changing $\text{rank}(\Phi|G)$. Therefore we obtain

Lemma 6.3 (Owren & Zennaro 1991). *Consider a Runge-Kutta method of order p . For the construction of a continuous extension of order $p^* = p$ one has to add at least*

$$\delta := \text{rank}(\Phi|G) - \text{rank}(\Phi) \quad (6.16)$$

stages. □

For the Dormand & Prince method we thus need at least two additional stages. There are several possibilities for constructing such dense output formulas:

- a) Shampine (1986) shows that one new function evaluation allows one to compute a 5th order approximation at the midpoint $x_0 + h/2$. If one evaluates anew the function at this point to get an approximation of $y'(x_0 + h/2)$, one can do quintic Hermite interpolation to get a dense output of order 5.

- b) Use the 4th order formula constructed above at two different output points and do boot-strapping. This has been done by Calvé & Vaillancourt (1990).
- c) Add two arbitrary new stages and solve the order conditions. This leads to methods with 10 free parameters (Calvo, Montijano & Rández 1992) which can then be used to minimize the error terms. This seems to give the best output formulas.

New methods. If anyhow the Dormand & Prince pair needs two additional function evaluations for a 5th order dense output, the suggestion lies at hand to search for completely new methods which use *all* stages for the solution y_1 and \hat{y}_1 as well. Owren & Zennaro (1992) constructed an 8-stage continuous Runge-Kutta method of order 5(4). It uses the FSAL idea so that the effective cost is 7 function evaluations (*fe*) per step. Bogacki & Shampine (1989) present a 7-stage method of order 5(4) with very small error coefficients, so that it nearly behaves like a 6th order method. The effective cost of its dense output is 10 *fe*. A method of order 6(5) with a dense output of order $p^* = 5$ is given by Calvo, Montijano & Rández (1990).

Dense Output for DOP853

We are interested in a continuous extension of the 8th order method of Section II.5 (formula (5.20)). A dense output of order 6 can be obtained for free (add y_1 as 13th stage and solve the linear system (6.19a-c) below with $s^* = s + 1 = 13$). Following Dormand & Prince we shall construct a dense output of order $p^* = 7$. We add three further stages (by Lemma 6.3 this is the minimal number of additional stages). The values for c_{14}, c_{15}, c_{16} are chosen arbitrarily as

$$c_{14} = 0.1, \quad c_{15} = 0.2, \quad c_{16} = 7/9 \quad (6.17)$$

and the coefficients a_{ij} are assumed to satisfy, for $i \in \{14, 15, 16\}$,

$$\sum_{j=1}^{i-1} a_{ij} c_j^{q-1} = c_i^q / q, \quad q = 1, \dots, 6 \quad (6.18a)$$

$$a_{i2} = a_{i3} = a_{i4} = a_{i5} = 0 \quad (6.18b)$$

$$\sum_{j=k+1}^{i-1} a_{ij} a_{jk} = 0, \quad k = 4, 5. \quad (6.18c)$$

This system can easily be solved (step 5 of Fig. 5.3). We are still free to set some coefficients equal to 0 (see Fig. 5.3).

We next search for polynomials $b_i(\theta)$ such that the conditions (6.4) are satisfied for all trees of order ≤ 7 . We find the following necessary conditions ($s^* = 16$)

$$\sum_{i=1}^{s^*} b_i(\theta) c_i^{q-1} = \theta^q / q, \quad q = 1, \dots, 7 \quad (6.19a)$$

$$b_2(\theta) = b_3(\theta) = b_4(\theta) = b_5(\theta) = 0 \quad (6.19b)$$

$$\sum_{i=j+1}^{s^*} b_i(\theta) a_{ij} = 0, \quad j = 4, 5 \quad (6.19c)$$

$$\sum_{i=j+1}^{s^*} b_i(\theta) c_i a_{ij} = 0, \quad j = 4, 5 \quad (6.19d)$$

$$\sum_{i,j=1}^{s^*} b_i(\theta) a_{ij} c_j^5 = \theta^7/42. \quad (6.19e)$$

Here (6.19a,e) are order conditions for $[\tau, \dots, \tau]$ and $[[\tau, \tau, \tau, \tau, \tau]]$. The property $b_2(\theta) = 0$ follows from $0 = \sum_i b_i(\theta) (\sum_j a_{ij} c_j - c_i^2/2) = -b_2(\theta) c_2^2/2$ and the other three conditions of (6.19b) are a consequence of the relations $0 = \sum_i b_i(\theta) c_i^{q-1} (\sum_j a_{ij} c_j^3 - c_i^4/4) = 0$ for $q = 1, 2, 3$. The necessity of the conditions (6.19c,d) is seen similarly.

On the other hand, the conditions (6.19) are also sufficient for the dense output to be of order 7. We first remark that (6.19), (6.18) and (5.20) imply

$$\sum_{i,j=k+1}^{s^*} b_i(\theta) a_{ij} a_{jk} = 0, \quad k = 4, 5 \quad (6.20)$$

(see Exercise 3). The verification of the order conditions (6.4) is then possible without difficulty.

System (6.19) consists of 16 linear equations for 16 unknowns which possess a unique solution. An interesting property of the continuous solution (6.1) obtained in this manner is that it yields a global \mathcal{C}^1 -approximation to the solution, i.e.,

$$u(0) = y_0, \quad u(1) = y_1, \quad u'(0) = hf(y_0), \quad u'(1) = hf(y_1). \quad (6.21)$$

For the verification of this property we define a polynomial $q(\theta)$ of degree 7 by the relations (6.21) and by $q(\theta_i) = u(\theta_i)$ for 4 distinct values θ_i which are different from 0 and 1. Obviously, $q(\theta)$ is of the form (6.1) and defines a dense output of order 7. Due to the uniqueness of the $b_i(\theta)$ we must have $q(\theta) \equiv u(\theta)$ so that (6.21) is verified.

Event Location

Often the output value x_{end} for which the solutions are wanted is not known in advance, but depends implicitly on the computed solutions. An example of such a situation is the search for periodic solutions and limit cycles discussed in Section I.16, where we wanted to know when the solution reaches the Poincaré-section for the first time.

Such problems are very easily treated when a dense output $u(x)$ is available. Suppose we want to determine x such that

$$g(x, y(x)) = 0. \quad (6.22)$$

Algorithm 6.4. Compute the solution step-by-step until a sign change appears between $g(x_i, y_i)$ and $g(x_{i+1}, y_{i+1})$ (this is, however, not completely safe because g may change sign twice in an integration interval; use the dense output at intermediate values if more safety is needed). Then replace $y(x)$ in (6.22) by the

approximation $u(x)$ and solve the resulting equation numerically, e.g. by bisection or Newton iterations.

This algorithm can be conveniently done in the subroutine SOLOUT, which is called after every accepted step (see Appendix). If the value of x , satisfying (6.22), has been found, the integration is stopped by setting ITRN = -1.

Whenever the function g of (6.22) also depends on $y'(x)$, it is advisable to use a dense output of order $p^* = p$.

Discontinuous Equations

If you write some software which is half-way useful, sooner or later someone will use it on discontinuities. You have to scope about ...
(A.R. Curtis 1986)

In many applications the function defining a differential equation is not analytic or continuous everywhere. A common example is a problem which (at least locally) can be written in the form

$$y' = \begin{cases} f_I(y) & \text{if } g(y) > 0 \\ f_{II}(y) & \text{if } g(y) < 0 \end{cases} \quad (6.23)$$

with sufficiently differentiable functions g , f_I and f_{II} . The derivative of the solution is thus in general discontinuous on the surface

$$S = \{y; g(y) = 0\}.$$

The function $g(y)$ is called a *switching function*.

In order to understand the situations which can occur when the solution of (6.23) meets the surface S in a point y_0 (i.e., $g(y_0) = 0$), we consider the scalar products

$$\begin{aligned} a_I &= \langle \text{grad } g(y_0), f_I(y_0) \rangle \\ a_{II} &= -\langle \text{grad } g(y_0), f_{II}(y_0) \rangle \end{aligned} \quad (6.24)$$

which can be approximated numerically by $a_I \approx g(y_0 + \delta f_I(y_0)) / \delta$ with small enough δ . Since the vector $\text{grad } g(y_0)$ points towards the domain of f_I , the inequality $a_I < 0$ tells us that the flow for f_I is “pushing” against S , while for $a_I > 0$ the flow is “pulling”. The same argument holds for a_{II} and the flow for f_{II} . Therefore, apart from degenerate cases where either a_I or a_{II} vanishes, we can distinguish the following four cases (see Fig. 6.2):

- 1) $a_I > 0, a_{II} < 0$: the flow traverses S from $g < 0$ to $g > 0$.
- 2) $a_I < 0, a_{II} > 0$: the flow traverses S from $g > 0$ to $g < 0$.
- 3) $a_I > 0, a_{II} > 0$: the flow “pulls” on both sides; the solution is not unique; except in the case of an unhappily chosen initial value, this situation would normally not occur.

- 4) $a_I < 0, a_{II} < 0$: here *both* flows push against S ; the solution is trapped in S and the problem no longer has a classical solution.

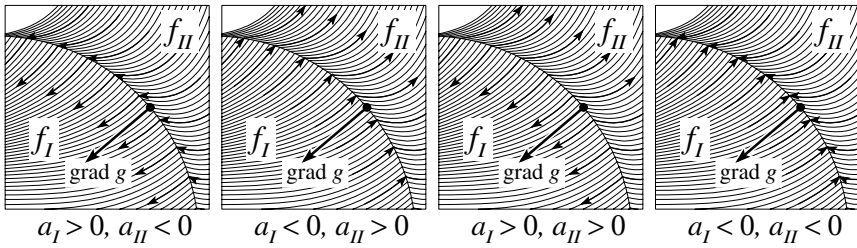


Fig. 6.2. Solutions near the surface of discontinuity

Crossing a discontinuity. The *numerical* computation of a solution crossing a discontinuity (cases 1 and 2) can be performed as follows:

- a) *Ignoring the discontinuity*: apply a variable step size code with local error control (such as DOPRI5) and hope that the step size mechanism would handle the discontinuity appropriately. Consider the example (which represents the flow of the second picture of Fig. 6.2)

$$y' = \begin{cases} x^2 + 2y^2 & \text{if } (x + 0.05)^2 + (y + 0.15)^2 \leq 1 \\ 2x^2 + 3y^2 - 2 & \text{if } (x + 0.05)^2 + (y + 0.15)^2 > 1 \end{cases} \quad (6.25)$$

with initial value $y(0) = 0.3$. The discontinuity for this problem occurs at $x \approx 0.6234$ and the code, applied with $Atol = Rtol = 10^{-5}$, detects the discontinuity fairly well by means of numerous rejected steps (see Fig. 6.3; this figure, however, is much less dramatic than an analogous drawing (see Gear & Østerby 1984) for multistep methods). The numerical solution for $x = 1$ then has an error of $5.9 \cdot 10^{-4}$.

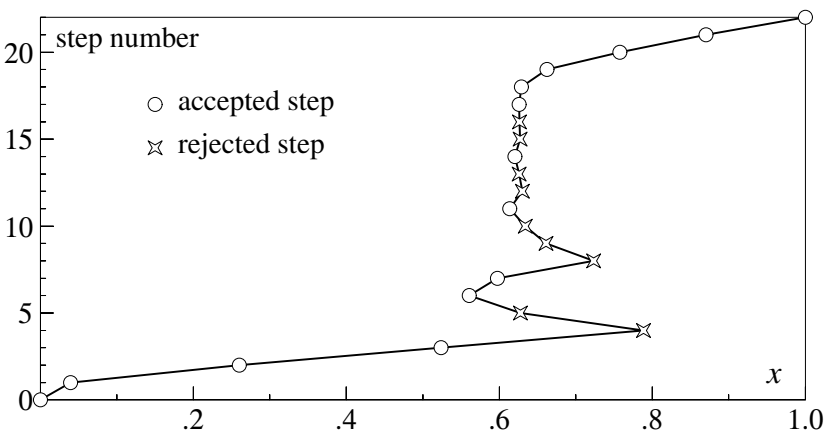


Fig. 6.3. Ignoring the discontinuity at problem (6.23)

- b) *Singularity detecting codes.* Concepts have been developed (Gear & Østerby (1984) for multistep methods, Enright, Jackson, Nørsett & Thomsen (1988) for Runge-Kutta methods) to modify existing codes in such a way that singularities are detected more precisely and handled more appropriately. These concepts are mainly based on the behaviour of the local error estimate compared to the step size.
- c) *Use the switching function:* stop the computation at the surface of discontinuity using Algorithm 6.4 and restart the integration with the new right-hand side. One has to take care that during one integration step only function values of either f_I or f_{II} are used. This algorithm, applied to Example (6.25), uses less than half of the function evaluations as the “ignoring algorithm” and gives an error of $6.6 \cdot 10^{-6}$ at the point $x = 1$. It is thus not only faster, but also much more reliable.

Example 6.5. Coulomb’s law of friction (Coulomb 1785), which states that the force of friction is *independent* of the speed, gives rise to many situations with discontinuous differential equations. Consider the example (see Den Hartog 1930, Reissig 1954, Taubert 1976)

$$y'' + 2Dy' + \mu \operatorname{sign} y' + y = A \cos(\omega x). \quad (6.26)$$

where the Coulomb-force $\mu \operatorname{sign} y'$ is accompanied by a viscosity term Dy' . We fix the parameters as $D = 0.1$, $\mu = 4$, $A = 2$ and $\omega = \pi$, and choose the initial values

$$y(0) = 3, \quad y'(0) = 4. \quad (6.27)$$

Equation (6.26), written in the form (6.23), is

$$\begin{aligned} y' &= v \\ v' &= -0.2v - y + 2 \cos(\pi x) - \begin{cases} 4 & \text{if } v > 0 \\ -4 & \text{if } v < 0. \end{cases} \end{aligned} \quad (6.28)$$

Its solution is plotted in Fig. 6.4.

The initial value (6.27) is in the region $v > 0$ and we follow the solution until it hits the manifold $v = 0$ for the first time. This happens for $x_1 \approx 0.5628$. An investigation of the values

$$a_I = -y(x_1) + 2 \cos(\pi x_1) - 4, \quad a_{II} = y(x_1) - 2 \cos(\pi x_1) - 4 \quad (6.29)$$

shows that $a_I < 0$, $a_{II} > 0$, so that we have to continue the integration into the region $v < 0$. The next intersection of the solution with the manifold of discontinuity is at $x_2 \approx 2.0352$. Here $a_I < 0$, $a_{II} < 0$, so that a classical solution does not exist beyond this point and the solution remains “trapped” in the manifold ($v = 0$, $y = \text{Const} = y(x_2)$) until one of the values a_I or a_{II} changes sign. This happens for a_{II} at the point $x_3 \approx 2.6281$ and we can continue the integration of (6.28) in the region $v < 0$ (see Fig. 6.4). The same situation then repeats periodically.

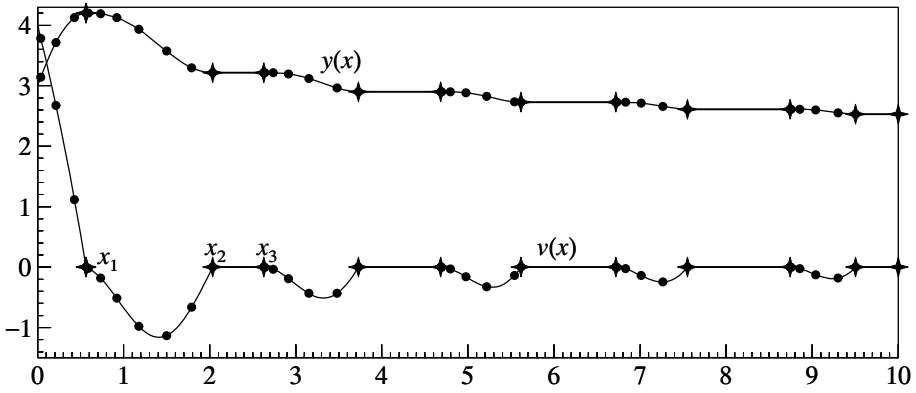


Fig. 6.4. Solutions of (6.28)

Solutions in the manifold. In the case $a_I < 0$, $a_{II} < 0$ the solution of (6.23) can neither be continued along the flow of $y' = f_I(y)$ nor along that of $y' = f_{II}(y)$. However, the physical process, described by the differential equation (6.23), possesses a solution (see Example 6.5). Early papers on this subject studied the convergence of Euler polygons, pushed across the border again and again by the conflicting vector fields (see, e.g., Taubert 1976). Later it became clear that it is much more advantageous to pursue the solution *in* the manifold S , i.e., solve a so-called differential algebraic problem. This approach is advocated by Eich (1992), who attributes the ideas to the thesis of G. Bock, by Eich, Kastner-Maresch & Reich (unpublished manuscript, 1991), and by Stewart (1990). We must decide, however, *which* vector field in S should determine the solution. Several motivations (see Exercises 8 and 9 below) suggest to search this field in the convex hull

$$f(y, \lambda) = (1 - \lambda)f_I(y) + \lambda f_{II}(y), \quad (6.30)$$

of f_I and f_{II} . This coincides, for the special problem (6.23), with Filippov's "generalized solution" (Filippov 1960); but other homotopies may be of interest as well. The value of λ must be chosen in such a way that the solution remains in S . This means that we have to solve the problem

$$y' = f(y, \lambda) \quad (6.31a)$$

$$0 = g(y). \quad (6.31b)$$

Differentiating (6.31b) with respect to time yields

$$0 = \text{grad } g(y)y' = \text{grad } g(y)f(y, \lambda). \quad (6.32)$$

If this relation allows λ to be expressed as a function of y , say as $\lambda = G(y)$, then (6.31a) becomes the ordinary differential equation

$$y' = f(y, G(y)) \quad (6.33)$$

which can be solved by standard integration methods. Obviously, the solution of

(6.33) together with $\lambda = G(y)$ satisfy (6.32) and after integration also (6.31b) (because the initial value satisfies $g(y_0) = 0$).

For the homotopy (6.30) the relation (6.32) becomes

$$(1 - \lambda)a_I(y) - \lambda a_{II}(y) = 0, \quad \text{i.e.,} \quad \lambda = \frac{a_I(y)}{a_I(y) + a_{II}(y)}, \quad (6.34)$$

where $a_I(y)$ and $a_{II}(y)$ are given in (6.24).

Remark . Problem (6.31) is a “differential-algebraic system of index 2” and direct numerical methods are discussed in Chapter VI of Volume II. The instances where a_I or a_{II} change sign can again be computed by using a dense output and Algorithm 6.4.

Numerical Computation of Derivatives with Respect to Initial Values and Parameters

For the efficient computation of boundary value problems by a shooting technique as explained in Section I.15, we need to compute the derivatives of the solutions with respect to (the missing) initial values. Also, if we want to adjust unknown parameters from given data, say by a nonlinear least squares procedure, we have to compute the derivatives of the solutions with respect to parameters in the differential equation.

We shall restrict our discussion to the problem

$$y' = f(x, y, B), \quad y(x_0) = y_0(B) \quad (6.35)$$

where the right-hand side function and the initial values depend on a real parameter B . The generalization to more than one parameter is straightforward. There are several possibilities for computing the derivative $\partial y / \partial B$.

External differentiation. Denote the numerical solution, obtained by a variable step size code with a fixed tolerance, by $y_{Tol}(x_{\text{end}}, x_0, B)$. Then the most simple device is to approximate the derivative by a finite difference

$$\frac{1}{\Delta B} \left(y_{Tol}(x_{\text{end}}, x_0, B + \Delta B) - y_{Tol}(x_{\text{end}}, x_0, B) \right). \quad (6.36)$$

However, due to the error control mechanism with its IF's and THEN's and step rejections, the function $y_{Tol}(x_{\text{end}}, x_0, B)$ is by no means a smooth function of the parameter B . Therefore, the errors of the two numerical results in (6.36) are not correlated, so that the error of (6.36) as an approximation to $\partial y / \partial B(x_{\text{end}}, x_0, B)$ is of size $\mathcal{O}(\text{Tot}/\Delta B) + \mathcal{O}(\Delta B)$, the second term coming from the discretization (6.36). This suggests taking for ΔB something like $\sqrt{\text{Tot}}$, and the error of (6.36) becomes of size $\mathcal{O}(\sqrt{\text{Tot}})$.

Internal differentiation. We know from Section I.14 that $\Psi = \partial y / \partial B$ is the solution of the variational equation

$$\Psi' = \frac{\partial f}{\partial y}(x, y, B)\Psi + \frac{\partial f}{\partial B}(x, y, B), \quad \Psi(x_0) = \frac{\partial y_0}{\partial B}(B). \quad (6.37)$$

Here y is the solution of (6.35). Hence, (6.35) and (6.37) together constitute a differential system for y and Ψ , which can be solved simultaneously by any code. If the partial derivatives $\partial f / \partial y$ and $\partial f / \partial B$ are available analytically, then the error of $\partial y / \partial B$, obtained by this procedure, is obviously of size Tol . This algorithm is equivalent to “internal differentiation” as introduced by Bock (1981).

If $\partial f / \partial y$ and $\partial f / \partial B$ are not available one can approximate them by finite differences so that (6.37) becomes

$$\Psi' = \frac{1}{\Delta B} \left(f(x, y + \Delta B \cdot \Psi, B + \Delta B) - f(x, y, B) \right). \quad (6.38)$$

The solution of (6.38), when inserted into (6.37), gives raise to a defect of size $\mathcal{O}(\Delta B) + \mathcal{O}(eps / \Delta B)$, where eps is the precision of the computer (independent of Tol). By Theorem I.10.2, the difference of the solutions of (6.38) and (6.37) is of the same size. Choosing $\Delta B \approx \sqrt{eps}$ the error of the approximation to $\partial y / \partial B$, obtained by solving (6.35), (6.38), will be of order $Tol + \sqrt{eps}$, so that for $Tol \geq \sqrt{eps}$ the result is as precise as that obtained by integration of (6.37). Observe that external differentiation and the numerical solution of (6.35), (6.38) need about the same number of function evaluations.

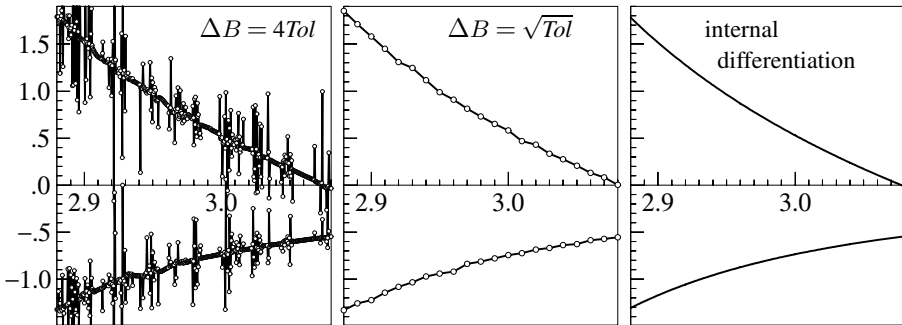


Fig. 6.5. Derivatives of the solution of (6.39) with respect to B

As an example we consider the Brusselator

$$\begin{aligned} y_1' &= 1 + y_1^2 y_2 - (B + 1)y_1 & y_1(0) &= 1.3 \\ y_2' &= B y_1 - y_1^2 y_2 & y_2(0) &= B \end{aligned} \quad (6.39)$$

and compute $\partial y / \partial B$ at $x = 20$ for various B ranging from $B = 2.88$ to $B = 3.08$. We applied the code DOPRI5 with $Atol = Rtol = Tol = 10^{-4}$. The numerical

result is displayed in Fig. 6.5. External differentiation has been applied, once with $\Delta B = \sqrt{\text{Tol}}$ and a second time with $\Delta B = 4\text{Tol}$. This numerical example clearly demonstrates that internal differentiation is to be preferred.

Exercises

1. (Owren & Zennaro 1991, Carnicer 1991). The 4-stage 4th order methods of Section II.1 do not possess a dense output of order 4 (also if the numerical solution y_1 is included as 5th stage). Prove this statement.
2. Consider a Runge-Kutta method of order p and use Richardson extrapolation for step size control. Besides the numerical solution y_0, y_1, y_2 we consider the extrapolated values (see Section II.4)

$$\hat{y}_1 = y_1 + \frac{y_2 - w}{(2^p - 1)2}, \quad \hat{y}_2 = y_2 + \frac{y_2 - w}{2^p - 1}$$

and do quintic polynomial interpolation based on $y_0, f(x_0, y_0), \hat{y}_1, f(x_0 + h, y_1), \hat{y}_2, f(x_0 + 2h, \hat{y}_2)$. Prove that the resulting dense output formula is of order $p^* = \min(5, p + 1)$.

Remark. It is not necessary to evaluate f at \hat{y}_1 .

3. Prove that the conditions (6.19), (6.18) and (5.20) imply (6.20).

Hint. The system (6.19) together with one relation of (6.20) is overdetermined. However, it possesses the solution b_i for $\theta = 1$. Further, the values $b_i c_i$ also solve this system if the right-hand side of (6.19a) is adapted. These properties imply that for $k \in \{4, 5\}$ and for $i \in \{1, 6, \dots, 16\}$

$$\sum_{j=k+1}^{i-1} a_{ij} a_{jk} = \alpha a_{i4} + \beta a_{i5} + \gamma c_i a_{i4} + \delta c_i a_{i5} + \varepsilon \left(\sum_{j=1}^{i-1} a_{ij} c_j^5 - \frac{c_i^6}{6} \right),$$

where the parameters $\alpha, \beta, \gamma, \delta, \varepsilon$ may depend on k .

4. (Butcher). Try your favorite code on the example

$$\begin{aligned} y_1' &= f_1(y_1, y_2), & y_1(0) &= 1 \\ y_2' &= f_2(y_1, y_2), & y_2(0) &= 0 \end{aligned}$$

where f is defined as follows.

If $(|y_1| > |y_2|)$ then

$$f_1 = 0, \quad f_2 = \text{sign}(y_1)$$

Else

$$f_2 = 0, \quad f_1 = -\text{sign}(y_2)$$

End If .

Compute $y_1(8), y_2(8)$. Show that the exact solution is periodic.

5. Do numerical computations for the problem $y' = f(y)$, $y(0) = 1$, $y(3) = ?$ where

$$f(y) = \begin{cases} y^2 & \text{if } 0 \leq y \leq 2 \\ \begin{matrix} \text{a) } 1 \\ \text{b) } 4 \\ \text{c) } -4 + 4y \end{matrix} & \text{if } 2 < y \end{cases}$$

Remark. The correct answer would be (a) 4.5, (b) 12, (c) $\exp(10) + 1$.

6. Consider an s -stage Runge-Kutta method and denote by \tilde{s} the number of distinct c_i . Prove that the order of any continuous extension is $\leq \tilde{s}$.

Hint. Let $q(x)$ be a polynomial of degree \tilde{s} satisfying $q(c_i) = 0$ (for $i = 1, \dots, s$) and investigate the expression $\sum_i b_i(\theta)q(c_i)$.

7. (Step size freeze). Consider the following algorithm for the computation of $\partial y / \partial B$: first compute numerically the solution of (6.35) and denote it by $y_h(x_{\text{end}}, B)$. At the same time memorize all the selected step sizes. This step size sequence is then used to solve (6.35) with B replaced by $B + \Delta B$. The result is denoted by $y_h(x_{\text{end}}, B + \Delta B)$. Then approximate the derivative $\partial y / \partial B$ by

$$\frac{1}{\Delta B} (y_h(x_{\text{end}}, B + \Delta B) - y_h(x_{\text{end}}, B)).$$

Prove that this algorithm is equivalent to the solution of the system (6.35), (6.38), if only the components of y are considered for error control and step size selection.

Remark. For large systems this algorithm needs less storage requirements than internal differentiation, in particular if the derivative with respect to several parameters is computed.

8. (Taubert 1976). Show that for the discontinuous problem (6.23) the Euler polygons converge to Filippov's solution (6.30), (6.31).

Hint. The difference quotient of a piece of the Euler polygon lies in the convex hull of points $f_I(y)$ and $f_{II}(y)$.

Remark. This result can either be interpreted as pleading for myriads of Euler steps, or as a motivation for the homotopy (6.30).

9. Another motivation for formula (6.30): suppose that a small particle of radius ε is transported in a possibly discontinuous flow. Then its movement might be described by the mean of f

$$f_\varepsilon(y) = \int_{B_\varepsilon(y)} f(z) dz / \int_{B_\varepsilon(y)} dz$$

which is continuous in y . Show that the solution of $y'_\varepsilon = f_\varepsilon(y)$ becomes, for $\varepsilon \rightarrow 0$, that of (6.33) and (6.34).

II.7 Implicit Runge-Kutta Methods

It has been traditional to consider only explicit processes
(J.C. Butcher 1964a)

The high speed computing machines make it possible to enjoy
the advantage of intricate methods
(P.C. Hammer & J.W. Hollingsworth 1955)

The first *implicit* RK methods were used by Cauchy (1824) for the sake of — you have guessed correctly — error estimation (Méthodes diverses qui peuvent être employées au Calcul numérique ...; see Exercise 5). Cauchy inserted the mean value theorem into the integral studied in Sections I.8 and II.1,

$$y(x_1) = y(x_0) + \int_{x_0}^{x_1} f(x, y(x)) dx, \quad (7.1)$$

to obtain

$$y_1 = y_0 + hf(x_0 + \theta h, y_0 + \Theta(y_1 - y_0)) \quad (7.2)$$

with $0 \leq \theta, \Theta \leq 1$ (the “ θ -method”). The extreme cases are $\theta = \Theta = 0$ (the explicit Euler method) and $\theta = \Theta = 1$

$$y_1 = y_0 + hf(x_1, y_1), \quad (7.3)$$

which we call the *implicit* or *backward Euler method*.

For the sake of more efficient numerical processes, we apply, as we did in Section II.1, the midpoint rule ($\theta = \Theta = 1/2$) and obtain from (7.2) by setting $k_1 = (y_1 - y_0)/h$:

$$\begin{aligned} k_1 &= f\left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2} k_1\right), \\ y_1 &= y_0 + hk_1. \end{aligned} \quad (7.4)$$

This method is called the *implicit midpoint rule*.

Still another possibility is to approximate (7.1) by the *trapezoidal rule* and to obtain

$$y_1 = y_0 + \frac{h}{2} \left(f(x_0, y_0) + f(x_1, y_1) \right). \quad (7.5)$$

Let us also look at the Radau scheme

$$\begin{aligned} y(x_1) - y(x_0) &= \int_{x_0}^{x_0+h} f(x, y(x)) dx \\ &\approx \frac{h}{4} \left(f(x_0, y_0) + 3f\left(x_0 + \frac{2}{3}h, y\left(x_0 + \frac{2}{3}h\right)\right) \right). \end{aligned}$$

Here we need to approximate $y(x_0 + 2h/3)$. One idea would be the use of quadratic interpolation based on y_0 , y'_0 and $y(x_1)$,

$$y\left(x_0 + \frac{2}{3}h\right) \approx \frac{5}{9}y_0 + \frac{4}{9}y(x_1) + \frac{2}{9}hf(x_0, y_0).$$

The resulting method, given by Hammer & Hollingsworth (1955), is

$$\begin{aligned} k_1 &= f(x_0, y_0) \\ k_2 &= f\left(x_0 + \frac{2}{3}h, y_0 + \frac{h}{3}(k_1 + k_2)\right) \\ y_1 &= y_0 + \frac{h}{4}(k_1 + 3k_2). \end{aligned} \quad (7.6)$$

All these schemes are of the form (1.8) if the summations are extended up to “ s ”.

Definition 7.1. Let b_i , a_{ij} ($i, j = 1, \dots, s$) be real numbers and let c_i be defined by (1.9). The method

$$\begin{aligned} k_i &= f\left(x_0 + c_i h, y_0 + h \sum_{j=1}^s a_{ij} k_j\right) \quad i = 1, \dots, s \\ y_1 &= y_0 + h \sum_{i=1}^s b_i k_i \end{aligned} \quad (7.7)$$

is called an *s-stage Runge-Kutta method*. When $a_{ij} = 0$ for $i \leq j$ we have an explicit (ERK) method. If $a_{ij} = 0$ for $i < j$ and at least one $a_{ii} \neq 0$, we have a *diagonal implicit Runge-Kutta method* (DIRK). If in addition all diagonal elements are identical ($a_{ii} = \gamma$ for $i = 1, \dots, s$), we speak of a *singly diagonal implicit* (SDIRK) method. In all other cases we speak of an *implicit Runge-Kutta method* (IRK).

The tableau of coefficients used above for ERK-methods is obviously extended to include all the other non-zero a_{ij} ’s above the diagonal. For methods (7.3), (7.4) and (7.6) it is given in Table 7.1.

Renewed interest in implicit Runge-Kutta methods arose in connection with *stiff* differential equations (see Volume II).

Table 7.1. Implicit Runge-Kutta methods

			0	0	0
			2/3	1/3	1/3
				1/4	3/4
1	1	1/2	1/2		
	1		1		
Implicit Euler	Implicit midpoint rule	Hammer & Hollingsworth			

Existence of a Numerical Solution

For implicit methods, the k_i 's can no longer be evaluated successively, since (7.7) constitutes a system of implicit equations for the determination of k_i . For DIRK-methods we have a sequence of implicit equations of dimension n for k_1 , then for k_2 , etc. For fully implicit methods $s \cdot n$ unknowns (k_i , $i = 1, \dots, s$; each of dimension n) have to be determined simultaneously, which still increases the difficulty. A natural question is therefore (the reason for which the original version of Butcher (1964a) was returned by the editors): do equations (7.7) possess a solution at all?

Theorem 7.2. *Let $f : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ be continuous and satisfy a Lipschitz condition with constant L (with respect to y). If*

$$h < \frac{1}{L \max_i \sum_j |a_{ij}|} \quad (7.8)$$

there exists a unique solution of (7.7), which can be obtained by iteration. If $f(x, y)$ is p times continuously differentiable, the functions k_i (as functions of h) are also in C^p .

Proof. We prove the existence by iteration (“... on la résoudra facilement par des approximations successives ...”, Cauchy 1824)

$$k_i^{(m+1)} = f\left(x_0 + c_i h, y_0 + h \sum_{j=1}^s a_{ij} k_j^{(m)}\right).$$

We define $K \in \mathbb{R}^{sn}$ as $K = (k_1, \dots, k_s)^T$ and use the norm $\|K\| = \max_i (\|k_i\|)$. Then (7.7) can be written as $K = F(K)$ where

$$F_i(K) = f\left(x_0 + c_i h, y_0 + h \sum_{j=1}^s a_{ij} k_j\right), \quad i = 1, \dots, s.$$

The Lipschitz condition and a repeated use of the triangle inequality then show that

$$\|F(K_1) - F(K_2)\| \leq hL \max_{i=1, \dots, s} \sum_{j=1}^s |a_{ij}| \cdot \|K_1 - K_2\|$$

which from (7.8) is a contraction. The contraction mapping principle then ensures the existence and uniqueness of the solution and the convergence of the fixed-point iteration.

The differentiability result is ensured by the Implicit Function Theorem of classical analysis: (7.7) is written as $\Phi(h, K) = K - F(K) = 0$. The matrix of partial derivatives $\partial\Phi/\partial K$ for $h = 0$ is the identity matrix and therefore the solution of $\Phi(h, K) = 0$, which for $h = 0$ is $k_i = f(x_0, y_0)$, is continuously differentiable in a neighbourhood of $h = 0$. \square

If the assumptions on f in Theorem 7.2 are only satisfied in a neighbourhood of the initial value, then further restrictions on h are needed in order that the argument of f remains in this neighbourhood. Uniqueness is then only of local nature.

The step size restriction (7.8) becomes useless for stiff problems (L large). We return to this question in Vol. II, Sections IV.8 and IV.14.

The definition of *order* is the same as for explicit methods and the order conditions are derived in precisely the same way as in Section II.2.

Example 7.3. Let us study implicit two-stage methods of order 3: the order conditions become (see Theorem 2.1)

$$\begin{aligned} b_1 + b_2 &= 1, & b_1 c_1 + b_2 c_2 &= \frac{1}{2}, & b_1 c_1^2 + b_2 c_2^2 &= \frac{1}{3} \\ b_1(a_{11}c_1 + a_{12}c_2) + b_2(a_{21}c_1 + a_{22}c_2) &= \frac{1}{6}. \end{aligned} \quad (7.9)$$

The first three equations imply the following orthogonality relation (from the theory of Gaussian integration):

$$\int_0^1 (x - c_1)(x - c_2) dx = 0, \quad \text{i.e., } c_2 = \frac{2 - 3c_1}{3 - 6c_1} \quad (c_1 \neq 1/2) \quad (7.10)$$

and

$$b_1 = \frac{c_2 - 1/2}{c_2 - c_1}, \quad b_2 = \frac{c_1 - 1/2}{c_1 - c_2}.$$

In the fourth equation we insert $a_{21} = c_2 - a_{22}$, $a_{11} = c_1 - a_{12}$ and consider a_{12} and c_1 as free parameters. This gives

$$a_{22} = \frac{1/6 - b_1 a_{12}(c_2 - c_1) - c_1/2}{b_2(c_2 - c_1)}. \quad (7.11)$$

For $a_{12} = 0$ we obtain a one-parameter family of DIRK-methods of order 3. An SDIRK-method is obtained if we still require $a_{11} = a_{22}$ (Nørsett 1974b, Crouzeix 1975, see Table 7.2). For order 4 we have 4 additional conditions, with only two free parameters left. Nevertheless there exists a unique solution (see Table 7.3).

Table 7.2. SDIRK method, order 3

γ	γ	0	$\gamma = \frac{3 \pm \sqrt{3}}{6}$
$1 - \gamma$	$1 - 2\gamma$	γ	
	$1/2$	$1/2$	

Table 7.3. Hammer & Hollingsworth, order 4

$\frac{1}{2} - \frac{\sqrt{3}}{6}$	$\frac{1}{4}$	$\frac{1}{4} - \frac{\sqrt{3}}{6}$
$\frac{1}{2} + \frac{\sqrt{3}}{6}$	$\frac{1}{4} + \frac{\sqrt{3}}{6}$	$\frac{1}{4}$
	$1/2$	$1/2$

The Methods of Kuntzmann and Butcher of Order $2s$

It is clear that formula (7.4) and the method of Table 7.3 extend the one-point and two-point Gaussian quadrature formulas, respectively. Kuntzmann (1961) (see Ceschino & Kuntzmann 1963, p. 106) and Butcher (1964a) then discovered that for all s there exist IRK-methods of order $2s$. The main tools of proof are the following *simplifying assumptions*

$$\begin{aligned}
 B(p) : \quad & \sum_{i=1}^s b_i c_i^{q-1} = \frac{1}{q} \quad q = 1, \dots, p, \\
 C(\eta) : \quad & \sum_{j=1}^s a_{ij} c_j^{q-1} = \frac{c_i^q}{q} \quad i = 1, \dots, s, \quad q = 1, \dots, \eta, \\
 D(\zeta) : \quad & \sum_{i=1}^s b_i c_i^{q-1} a_{ij} = \frac{b_j}{q} (1 - c_j^q) \quad j = 1, \dots, s, \quad q = 1, \dots, \zeta.
 \end{aligned}$$

Condition $B(p)$ simply means that the quadrature formula (b_i, c_i) is of order p or, equivalently, that the order conditions (2.21) are satisfied for the bushy trees $[\tau, \dots, \tau]$ up to order p .

The assumption $C(\eta)$ implies that the pairs of trees in Fig. 7.1 give identical order conditions for $q \leq \eta$. In contrast to explicit Runge-Kutta methods (see (5.7) and (5.15)) there is no need to require conditions such as $b_2 = 0$ (see (5.8)), because $\sum_j a_{ij} c_j^{q-1} = c_i^q / q$ is valid for all i .

The assumption $D(\zeta)$ is an extension of (1.12). It means that the order condition of the left-hand tree of Fig. 7.2 is implied by those of the two right-hand trees if $q \leq \zeta$.

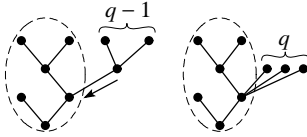


Fig. 7.1. Reduction with $C(q)$

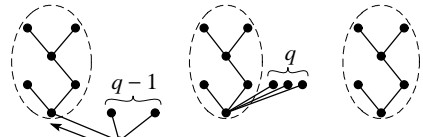


Fig. 7.2. Reduction with $D(q)$

Theorem 7.4 (Butcher 1964a). *If $B(p)$, $C(\eta)$ and $D(\zeta)$ are satisfied with $p \leq 2\eta + 2$ and $p \leq \zeta + \eta + 1$, then the method is of order p .*

Proof. The above reduction by $C(\eta)$ implies that it is sufficient to consider trees $t = [t_1, \dots, t_m]$ of order $\leq p$, where the subtrees t_1, \dots, t_m are either equal to τ or of order $\geq \eta + 1$. Since $p \leq 2\eta + 2$ either all subtrees are equal to τ or there is exactly one subtree different from τ . In the second case the number of τ 's is $\leq \zeta - 1$ by $p \leq \eta + \zeta + 1$ and the reduction by $D(\zeta)$ can be applied. Therefore, after all these reductions, only the bushy trees are left and they are satisfied by $B(p)$. \square

To obtain the formulas of order $2s$, Butcher assumed $B(2s)$ (i.e., the c_i and b_i are the coefficients of the Gaussian quadrature formula) and $C(s)$. This implies $D(s)$ (see Exercise 7) so that Theorem 7.4 can be applied with $p = 2s$, $\eta = s$ and $\zeta = s$. Hence the method, obtained in this way, is of order $2s$. For $s = 3$ and 4 the coefficients are given in Tables 7.4 and 7.5. They can still be expressed by radicals for $s = 5$ and are given in Butcher (1964a), p. 57.

Impressive numerical results from celestial mechanics for these methods were first reported in the thesis of D. Sommer (see Sommer 1965).

Table 7.4. Kuntzmann & Butcher method, order 6

$\frac{1}{2} - \frac{\sqrt{15}}{10}$	$\frac{5}{36}$	$\frac{2}{9} - \frac{\sqrt{15}}{15}$	$\frac{5}{36} - \frac{\sqrt{15}}{30}$
$\frac{1}{2}$	$\frac{5}{36} + \frac{\sqrt{15}}{24}$	$\frac{2}{9}$	$\frac{5}{36} - \frac{\sqrt{15}}{24}$
$\frac{1}{2} + \frac{\sqrt{15}}{10}$	$\frac{5}{36} + \frac{\sqrt{15}}{30}$	$\frac{2}{9} + \frac{\sqrt{15}}{15}$	$\frac{5}{36}$
	$\frac{5}{18}$	$\frac{4}{9}$	$\frac{5}{18}$

Table 7.5. Kuntzmann & Butcher method, order 8

$\frac{1}{2} - \omega_2$	ω_1	$\omega'_1 - \omega_3 + \omega'_4$	$\omega'_1 - \omega_3 - \omega'_4$	$\omega_1 - \omega_5$
$\frac{1}{2} - \omega'_2$	$\omega_1 - \omega'_3 + \omega_4$	ω'_1	$\omega'_1 - \omega'_5$	$\omega_1 - \omega'_3 - \omega_4$
$\frac{1}{2} + \omega'_2$	$\omega_1 + \omega'_3 + \omega_4$	$\omega'_1 + \omega'_5$	ω'_1	$\omega_1 + \omega'_3 - \omega_4$
$\frac{1}{2} + \omega_2$	$\omega_1 + \omega_5$	$\omega'_1 + \omega_3 + \omega'_4$	$\omega'_1 + \omega_3 - \omega'_4$	ω_1
	$2\omega_1$	$2\omega'_1$	$2\omega'_1$	$2\omega_1$
$\omega_1 = \frac{1}{8} - \frac{\sqrt{30}}{144},$	$\omega'_1 = \frac{1}{8} + \frac{\sqrt{30}}{144},$			
$\omega_2 = \frac{1}{2} \sqrt{\frac{15 + 2\sqrt{30}}{35}},$	$\omega'_2 = \frac{1}{2} \sqrt{\frac{15 - 2\sqrt{30}}{35}},$			
$\omega_3 = \omega_2 \left(\frac{1}{6} + \frac{\sqrt{30}}{24} \right),$	$\omega'_3 = \omega'_2 \left(\frac{1}{6} - \frac{\sqrt{30}}{24} \right),$			
$\omega_4 = \omega_2 \left(\frac{1}{21} + \frac{5\sqrt{30}}{168} \right),$	$\omega'_4 = \omega'_2 \left(\frac{1}{21} - \frac{5\sqrt{30}}{168} \right),$			
$\omega_5 = \omega_2 - 2\omega_3,$	$\omega'_5 = \omega'_2 - 2\omega'_3.$			

An important interpretation of the assumption $C(\eta)$ is the following:

Lemma 7.5. *The assumption $C(\eta)$ implies that the internal stages*

$$g_i = y_0 + h \sum_{j=1}^s a_{ij} k_j, \quad k_j = f(x_0 + c_j h, g_j) \quad (7.12)$$

satisfy for $i = 1, \dots, s$

$$g_i - y(x_0 + c_i h) = \mathcal{O}(h^{\eta+1}). \quad (7.13)$$

Proof. Because of $C(\eta)$ the exact solution satisfies (Taylor expansion)

$$y(x_0 + c_i h) = y_0 + h \sum_{j=1}^s a_{ij} y'(x_0 + c_j h) + \mathcal{O}(h^{\eta+1}). \quad (7.14)$$

Subtracting (7.14) from (7.12) yields

$$\begin{aligned} g_i - y(x_0 + c_i h) &= h \sum_{j=1}^s a_{ij} \left(f(x_0 + c_j h, g_j) - f(x_0 + c_j h, y(x_0 + c_j h)) \right) \\ &\quad + \mathcal{O}(h^{\eta+1}) \end{aligned}$$

and Lipschitz continuity of f proves (7.13). \square

IRK Methods Based on Lobatto Quadrature

Lobatto quadrature rules (Lobatto 1852, Radau 1880, p. 307) modify the idea of Gaussian quadrature by requiring that the first and the last node coincide with the interval ends, i.e., $c_1 = 0$, $c_s = 1$. These points are easier to handle and, in a step-by-step procedure, can be used twice. The remaining c 's are then adjusted optimally, i.e., as the zeros of the Jacobi orthogonal polynomial $P_{s-2}^{(1,1)}(x)$ or of $P'_{s-1}(x)$ (see e.g., Abramowitz & Stegun 1964, 25.4.32 for the interval $[-1, 1]$) and lead to formulas of order $2s - 2$.

J.C. Butcher (1964a, p. 51, 1964c) then found that Lobatto quadrature rules can be extended to IRK-methods whose coefficient matrix is zero in the first line and the last column. The first and the last stage then become *explicit* and the number of implicit stages reduces to $s - 2$. The methods are characterized by $B(2s - 2)$ and $C(s - 1)$. As in Exercise 7 this implies $D(s - 1)$ so that by Theorem 7.4 the method is of order $2s - 2$. For $s = 3$ and 4, the coefficients are given in Table 7.6.

We shall see in Volume II (Section IV.3, Table 3.1) that these methods, although preferable as concerns the relation between order and implicit stages, are not sufficiently stable for stiff differential equations.

Table 7.6. Butcher's Lobatto formulas of orders 4 and 6

				0	0	0	0	0
0	0	0	0	$\frac{5-\sqrt{5}}{10}$	$\frac{5+\sqrt{5}}{60}$	$\frac{1}{6}$	$\frac{15-7\sqrt{5}}{60}$	0
$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	0	$\frac{5+\sqrt{5}}{10}$	$\frac{5-\sqrt{5}}{60}$	$\frac{15+7\sqrt{5}}{60}$	$\frac{1}{6}$	0
1	0	1	0	1	$\frac{1}{6}$	$\frac{5-\sqrt{5}}{12}$	$\frac{5+\sqrt{5}}{12}$	0
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$		$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$

Collocation Methods

Es ist erstaunlich dass die Methode trotz ihrer Primitivität und der geringen Rechenarbeit in vielen Fällen ... sogar gute Ergebnisse liefert.
(L. Collatz 1951)

Nous allons montrer l'équivalence de notre définition avec la définition traditionnelle de certaines formules de Runge Kutta implicites.
(Guillou & Soulé 1969)

The concept of collocation is old and universal in numerical analysis (see e.g., pp. 28,29,32,181,411,453,483,495 of Collatz 1960, Frazer, Jones & Skan 1937). For ordinary differential equations it consists in searching for a polynomial of degree s whose derivative coincides ("co-locates") at s given points with the vector field of the differential equation (Guillou & Soulé 1969, Wright 1970). Still another approach is to combine Galerkin's method with numerical quadrature (see Hulme 1972).

Definition 7.6. For s a positive integer and c_1, \dots, c_s distinct real numbers (typically between 0 and 1), the corresponding *collocation polynomial* $u(x)$ of degree s is defined by

$$u(x_0) = y_0 \quad (\text{initial value}) \quad (7.15a)$$

$$u'(x_0 + c_i h) = f(x_0 + c_i h, u(x_0 + c_i h)), \quad i = 1, \dots, s. \quad (7.15b)$$

The numerical solution is then given by

$$y_1 = u(x_0 + h). \quad (7.15c)$$

If some of the c_i coincide, the collocation condition (7.15b) will contain higher derivatives and lead to multi-derivative methods (see Section II.13). Accordingly, for the moment, we suppose them all distinct.

Theorem 7.7 (Guillou & Soulé 1969, Wright 1970). *The collocation method (7.15) is equivalent to the s -stage IRK-method (7.7) with coefficients*

$$a_{ij} = \int_0^{c_i} \ell_j(t) dt, \quad b_j = \int_0^1 \ell_j(t) dt \quad i, j = 1, \dots, s, \quad (7.16)$$

where the $\ell_j(t)$ are the Lagrange polynomials

$$\ell_j(t) = \prod_{k \neq j} \frac{(t - c_k)}{(c_j - c_k)}. \quad (7.17)$$

Proof. Put $u'(x_0 + c_i h) = k_i$, so that

$$u'(x_0 + th) = \sum_{j=1}^s k_j \cdot \ell_j(t) \quad (\text{Lagrange}).$$

Then integrate

$$u(x_0 + c_i h) = y_0 + h \int_0^{c_i} u'(x_0 + th) dt \quad (7.18)$$

and insert into (7.15b) together with (7.16). The IRK-method (7.7) then comes out. \square

As a consequence of this result, the existence and uniqueness of the collocation polynomial (for sufficiently small h) follows from Theorem 7.2.

Theorem 7.8. *An implicit Runge-Kutta method with all c_i different and of order at least s is a collocation method iff $C(s)$ is true.*

Proof. $C(s)$ determines the a_{ij} uniquely. We write it as

$$\sum_{j=1}^s a_{ij} p(c_j) = \int_0^{c_i} p(t) dt \quad (7.19)$$

for all polynomials p of degree $\leq s - 1$. The a_{ij} given by (7.16) satisfy this relation, because (7.16) inserted into (7.19) is just the Lagrange interpolation formula. \square

Theorem 7.9. *Let $M(t) = \prod_{i=1}^s (t - c_i)$ and suppose that M is orthogonal to polynomials of degree $r - 1$,*

$$\int_0^1 M(t) t^{q-1} dt = 0, \quad q = 1, \dots, r, \quad (7.20)$$

then method (7.15) has order $p = s + r$.

Proof. The following proof uses the Gröbner & Alekseev Formula, which gives nice insight in the background of the result. An alternative proof is indicated in Exercise 7 below. One can also linearize the equation, apply the *linear* variation-of-constants formula and estimate the error (Guillou & Soulé 1969).

The orthogonality condition (7.20) means that the quadrature formula

$$\int_{x_0}^{x_0+h} g(t) dt = h \sum_{j=1}^s b_j g(x_0 + c_j h) + \text{err}(g) \quad (7.21)$$

is of order $s + r = p$, and its error is bounded by

$$|\text{err}(g)| \leq Ch^{p+1} \cdot \max |g^{(p)}(x)|. \quad (7.22)$$

The principal idea of the proof is now the following: we consider

$$u'(x) = f(x, u(x)) + (u'(x) - f(x, u(x)))$$

as a perturbation of

$$y'(x) = f(x, y(x))$$

and integrate the Gröbner & Alekseev Formula (I.14.18) with the quadrature formula (7.21). Due to (7.15b), the result is identically zero, since at the collocation points the defect is zero. Thus from (7.21) and (7.22)

$$\|y(x_0 + h) - u(x_0 + h)\| = \|\text{err}(g)\| \leq C \cdot h^{p+1} \cdot \max_{x_0 \leq t \leq x_0+h} \|g^{(p)}(t)\|, \quad (7.23)$$

where

$$g(t) = \frac{\partial y}{\partial y_0}(x, t, u(t)) \cdot (u'(t) - f(t, u(t))),$$

and we see that the local error behaves like $\mathcal{O}(h^{p+1})$.

There remains, however, a small technical detail: to show that the derivatives of $g(t)$ remain bounded for $h \rightarrow 0$. These derivatives contain partial derivatives of $f(t, y)$ and derivatives of $u(t)$. We shall see in the next theorem that these derivatives remain bounded for $h \rightarrow 0$. \square

Theorem 7.10. *The collocation polynomial $u(x)$ gives rise to a continuous IRK method of order s , i.e., for all $x_0 \leq x \leq x_0 + h$ we have*

$$\|y(x) - u(x)\| \leq C \cdot h^{s+1}. \quad (7.24)$$

Moreover, for the derivatives of $u(x)$ we have

$$\|y^{(k)}(x) - u^{(k)}(x)\| \leq C \cdot h^{s+1-k} \quad k = 0, \dots, s. \quad (7.25)$$

Proof. The exact solution $y(x)$ satisfies the collocation condition everywhere, hence *also* at the points $x_0 + c_i h$. So, in exactly the same way as in the proof

of Theorem 7.7, we apply the Lagrange interpolation formula to $y'(x)$:

$$y'(x_0 + th) = \sum_{j=1}^s f(x_0 + c_j h, y(x_0 + c_j h)) \ell_j(t) + h^s R(t, h)$$

where $R(t, h)$ is a smooth function of both variables. Integration and subtraction from (7.18) gives

$$y(x_0 + th) - u(x_0 + th) = h \sum_{j=1}^s \Delta f_j \cdot \int_0^t \ell_j(\tau) d\tau + h^{s+1} \int_0^t R(\tau, h) d\tau, \quad (7.26)$$

where

$$\Delta f_j = f(x_0 + c_j h, y(x_0 + c_j h)) - f(x_0 + c_j h, u(x_0 + c_j h)).$$

The k th derivative of (7.26) with respect to t is

$$h^k \left(y^{(k)}(x_0 + th) - u^{(k)}(x_0 + th) \right) = h \sum_{j=1}^s \Delta f_j \cdot \ell_j^{(k-1)}(t) + h^{s+1} \frac{\partial^{k-1} R}{\partial t^{k-1}}(t, h),$$

so that the result follows from the boundedness of the derivatives of $R(t, h)$ and from $\Delta f_j = \mathcal{O}(h^{s+1})$ which is a consequence of Lemma 7.5. \square

Remark. Only *some* IRK methods are collocation methods. An extension of the collocation idea (“Perturbed Collocation”, see Nørsett & Wanner 1981) applies to *all* IRK methods.

Exercises

1. Compute the one-point collocation method ($s = 1$) with $c_i = \theta$ and compare with (7.2). Determine its order in dependence of θ .
2. Compute all collocation methods with $s = 2$ of order 2 in dependence of c_1 and c_2 .
3. Specify in the method of Exercise 2 $c_1 = 1/3$, $c_2 = 1$ as well as $c_1 = 0$, $c_2 = 2/3$. Determine the orders of the obtained methods and explain.
4. Interpret the implicit midpoint rule (7.4) and the explicit Euler method as collocation methods. Is method (7.5) a collocation method? Method (7.6)?
5. (Cauchy 1824). Find from equation (7.2) conditions for the function $f(x, y)$ such that for scalar differential equations

$$y_1(\text{explicit Euler}) \geq y(x_1) \geq y_1(\text{implicit Euler}).$$

Compute five steps with $h = 0.2$ with both methods to obtain upper and lower bounds for $y(1)$, the solution of

$$y' = \cos \frac{x+y}{5}, \quad y(0) = 0.$$

Cauchy's result: $0.9659 \leq y(1) \leq 0.9810$. For one single step with $h = 1$ he obtained $0.926 \leq y(1) \leq 1$.

Compute the exact solution by elementary integration.

6. Determine the orders of the methods of Table 7.7. Generalize to arbitrary s (Ehle 1968).

Hint. Use Theorems 7.8 and 7.9.

Table 7.7. Methods of Ehle

Radau IIA, order 5				Lobatto IIIA, order 4			
$\frac{4-\sqrt{6}}{10}$	$\frac{88-7\sqrt{6}}{360}$	$\frac{296-169\sqrt{6}}{1800}$	$\frac{-2+3\sqrt{6}}{225}$	0	0	0	0
$\frac{4+\sqrt{6}}{10}$	$\frac{296+169\sqrt{6}}{1800}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{-2-3\sqrt{6}}{225}$	$\frac{1}{2}$	$\frac{5}{24}$	$\frac{1}{3}$	$-\frac{1}{24}$
1	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$	1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$		$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

7. (Butcher 1964a). Give an algebraic proof of Theorem 7.9.

Hint. From Theorem 7.8 we have $C(s)$.

Next the condition $B(p)$ with $p = s + r$ (theory of Gaussian quadrature formulas) implies $D(r)$. To see this, multiply the two vectors $u_j = \sum_i b_i c_i^{q-1} a_{ij}$ and $v_j = b_j(1 - c_j^q)/q$ ($j = 1, \dots, s$) by the Vandermonde matrix

$$V = \begin{pmatrix} 1 & 1 & \dots & 1 \\ c_1 & c_2 & \dots & c_s \\ \vdots & \vdots & \dots & \vdots \\ c_1^{s-1} & c_2^{s-1} & \dots & c_s^{s-1} \end{pmatrix}.$$

Finally apply Theorem 7.4.

II.8 Asymptotic Expansion of the Global Error

Mein Verzicht auf das Restglied war leichtsinnig . . .
(W. Romberg 1979)

Our next goal will be to perfect Richardson's extrapolation method (see Section II.4) by doing *repeated* extrapolation and eliminating more and more terms Ch^{p+k} of the error. A sound theoretical basis for this procedure is given by the study of the asymptotic behaviour of the global error. For problems of the type $y' = f(x)$, which lead to integration, the answer is given by the Euler-Maclaurin formula and has been exploited by Romberg (1955) and his successors. The first rigorous treatments for differential equations are due to Henrici (1962) and Gragg (1964) (see also Stetter 1973). We shall follow here the successive elimination of the error terms given by Hairer & Lubich (1984), which also generalizes to multistep methods.

Suppose we have a one-step method which we write, in Henrici's notation, as

$$y_{n+1} = y_n + h\Phi(x_n, y_n, h). \quad (8.1)$$

If the method is of order p , it possesses at each point of the solution $y(x)$ a *local error* of the form

$$\begin{aligned} y(x+h) - y(x) - h\Phi(x, y(x), h) = \\ d_{p+1}(x)h^{p+1} + \dots + d_{N+1}(x)h^{N+1} + \mathcal{O}(h^{N+2}) \end{aligned} \quad (8.2)$$

whenever the differential equation is sufficiently differentiable. For Runge-Kutta methods these error terms were computed in Section II.2 (see also Theorem 3.2).

The Global Error

Let us now set $y_n =: y_h(x)$ for the numerical solution at $x = x_0 + nh$. We then know from Theorem 3.6 that the global error behaves like h^p . We shall search for a function $e_p(x)$ such that

$$y(x) - y_h(x) = e_p(x)h^p + o(h^p). \quad (8.3)$$

The idea is to consider

$$y_h(x) + e_p(x)h^p =: \hat{y}_h(x) \quad (8.4a)$$

as the numerical solution of a new method

$$\widehat{y}_{n+1} = \widehat{y}_n + h\widehat{\Phi}(x_n, \widehat{y}_n, h). \quad (8.4b)$$

By comparison with (8.1), we see that the increment function for the new method is

$$\widehat{\Phi}(x, \widehat{y}, h) = \Phi(x, \widehat{y} - e_p(x)h^p, h) + (e_p(x+h) - e_p(x))h^{p-1}. \quad (8.5)$$

Our task is to find a function $e_p(x)$, with $e_p(x_0) = 0$, such that the method with increment function $\widehat{\Phi}$ is of order $p+1$.

Expanding the local error of the one-step method $\widehat{\Phi}$ into powers of h we obtain

$$\begin{aligned} y(x+h) - y(x) - h\widehat{\Phi}(x, y(x), h) \\ = \left(d_{p+1}(x) + \frac{\partial f}{\partial y}(x, y(x))e_p(x) - e'_p(x) \right) h^{p+1} + \mathcal{O}(h^{p+2}) \end{aligned} \quad (8.6)$$

where we have used

$$\frac{\partial \Phi}{\partial y}(x, y, 0) = \frac{\partial f}{\partial y}(x, y). \quad (8.7)$$

The term in h^{p+1} vanishes if $e_p(x)$ is defined as the solution of

$$e'_p(x) = \frac{\partial f}{\partial y}(x, y(x))e_p(x) + d_{p+1}(x), \quad e_p(x_0) = 0. \quad (8.8)$$

By Theorem 3.6, applied to the method $\widehat{\Phi}$, we now have

$$y(x) - y_h(x) = e_p(x)h^p + \mathcal{O}(h^{p+1}) \quad (8.9)$$

and the first term of the desired asymptotic expansion has been determined.

We now repeat the procedure with the method with increment function $\widehat{\Phi}$. It is of order $p+1$ and again satisfies condition (8.7). The final result of this procedure is the following

Theorem 8.1 (Gragg 1964). *Suppose that a given method with sufficiently smooth increment function Φ satisfies the consistency condition $\Phi(x, y, 0) = f(x, y)$ and possesses an expansion (8.2) for the local error. Then the global error has an asymptotic expansion of the form*

$$y(x) - y_h(x) = e_p(x)h^p + \dots + e_N(x)h^N + E_h(x)h^{N+1} \quad (8.10)$$

where the $e_j(x)$ are solutions of inhomogeneous differential equations of the form (8.8) with $e_j(x_0) = 0$ and $E_h(x)$ is bounded for $x_0 \leq x \leq x_{\text{end}}$ and $0 \leq h \leq h_0$. \square

The differentiability properties of the $e_j(x)$ depend on those of f and Φ (see (8.8) and (8.2)). The expansion (8.10) will be the theoretical basis for all discussions of extrapolation methods.

Examples. 1. For the equation $y' = y$ and Euler's method we have with $h = 1/n$ and $x = 1$, using the binomial theorem,

$$y_h(1) = \left(1 + \frac{1}{n}\right)^n = 1 + 1 + \left(1 - \frac{1}{n}\right) \frac{1}{2!} + \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \frac{1}{3!} + \dots$$

By multiplying out, this gives

$$y(1) - y_h(1) = - \sum_{i=1}^{\infty} h^i \sum_{j=1}^{\infty} \frac{S_{i+j}^{(j)}}{(i+j)!} = 1.359h - 1.246h^2 \pm \dots$$

where the $S_i^{(j)}$ are the Stirling numbers of the first kind (1730, see Abramowitz & Stegun 1964, Section 24.1.3). This is, of course, the Taylor series for the function

$$e - (1+h)^{1/h} = e - \exp\left(1 - \frac{h}{2} + \frac{h^2}{3} \pm \dots\right) = e\left(\frac{1}{2}h - \frac{11}{24}h^2 + \frac{7}{16}h^3 \pm \dots\right)$$

with *convergence radius* $r = 1$.

2. For the differential equation $y' = f(x)$ and the trapezoidal rule (7.5), the expansion (8.10) becomes

$$\int_0^1 f(x) dx - y_h(1) = - \sum_{k=1}^N \frac{h^{2k}}{(2k)!} B_{2k} \left(f^{(2k-1)}(1) - f^{(2k-1)}(0) \right) + \mathcal{O}(h^{2N+1}),$$

the well known Euler-Maclaurin formula (1736). For $N \rightarrow \infty$, the series will usually diverge, due to the fast growth of the Bernoulli numbers for large k . It may, however, be useful for small values of N and we call it an *asymptotic expansion* (Poincaré 1893).

Variable h

Theorem 8.1 is not only valid for equal step sizes. A reasonable assumption for the case of variable step sizes is the existence of a function $\tau(x) > 0$ such that the step sizes depend as

$$x_{n+1} - x_n = \tau(x_n) h \quad (8.11)$$

on a parameter h . Then the local error expansion (8.2) becomes

$$y(x + \tau(x)h) - y(x) - h\tau(x)\Phi(x, y(x), \tau(x)h) = d_{p+1}(x)\tau^{p+1}(x)h^{p+1} + \dots$$

and instead of (8.5) we have

$$\widehat{\Phi}(x, \widehat{y}, \tau(x)h) = \Phi(x, \widehat{y} - e_p(x)h^p, \tau(x)h) + \frac{h^p}{h\tau(x)} \left(e_p(x + \tau(x)h) - e_p(x) \right).$$

With this the local error expansion for the new method becomes, instead of (8.6),

$$y(x + \tau(x)h) - y(x) - h\tau(x)\widehat{\Phi}(x, y(x), \tau(x)h)$$

$$= \tau(x) \left(d_{p+1}(x) \tau^p(x) + \frac{\partial f}{\partial y}(x, y(x)) e_p(x) - e'_p(x) \right) h^{p+1} + \mathcal{O}(h^{p+2})$$

and the proof of Theorem 8.1 generalizes with slight modifications.

Negative h

The most important extrapolation algorithms will use asymptotic expansions with *even* powers of h . In order to provide a theoretical basis for these methods, we need to explain the meaning of $y_h(x)$ for h *negative*.

Motivation. We write (8.1) as

$$y_h(x+h) = y_h(x) + h\Phi(x, y_h(x), h) \quad (8.1')$$

and replace h by $-h$ to obtain

$$y_{-h}(x-h) = y_{-h}(x) - h\Phi(x, y_{-h}(x), -h).$$

Next we replace x by $x+h$ which gives

$$y_{-h}(x) = y_{-h}(x+h) - h\Phi(x+h, y_{-h}(x+h), -h). \quad (8.12)$$

This is an implicit equation for $y_{-h}(x+h)$, which possesses a unique solution for sufficiently small h (by the implicit function theorem). We write this solution in the form

$$y_{-h}(x+h) = y_{-h}(x) + h\Phi^*(x, y_{-h}(x), h). \quad (8.13)$$

The comparison of (8.12) and (8.13) (with $A = y_{-h}(x+h)$, $B = y_{-h}(x)$) leads us to the following definition.

Definition 8.2. Let $\Phi(x, y, h)$ be the increment function of a method. Then we define the increment function $\Phi^*(x, y, h)$ of the *adjoint method* by the pair of formulas

$$\begin{aligned} B &= A - h\Phi(x+h, A, -h) \\ A &= B + h\Phi^*(x, B, h). \end{aligned} \quad (8.14)$$

Example. The adjoint method of explicit Euler is implicit Euler.

Theorem 8.3. Let Φ be the Runge-Kutta method (7.7) with coefficients a_{ij} , b_j , c_i ($i, j = 1, \dots, s$). Then the adjoint method Φ^* is equivalent to a Runge-Kutta method with s stages and with coefficients

$$\begin{aligned} c_i^* &= 1 - c_{s+1-i} \\ a_{ij}^* &= b_{s+1-j} - a_{s+1-i, s+1-j} \\ b_j^* &= b_{s+1-j}. \end{aligned}$$

Proof. The formulas (8.14) indicate that for the definition of the adjoint method we have, starting from (7.7), to exchange $y_0 \leftrightarrow y_1$, $h \leftrightarrow -h$ and replace $x_0 \rightarrow x_0 + h$. This then leads to

$$k_i = f\left(x_0 + (1 - c_i)h, y_0 + h \sum_{j=1}^s (b_j - a_{ij})k_j\right)$$

$$y_1 = y_0 + h \sum_{j=1}^s b_j k_j.$$

In order to preserve the usual natural ordering of c_1, \dots, c_s , we also permute the k_i -values and replace all indices i by $s + 1 - i$. \square

Properties of the Adjoint Method

Theorem 8.4. $\Phi^{**} = \Phi$.

Proof. This property, which is the reason for the name “adjoint”, is seen by replacing $h \rightarrow -h$ and then $x \rightarrow x + h$, $B \rightarrow A$, $A \rightarrow B$ in (8.14). \square

Theorem 8.5. *The adjoint method has the same order as the original method. Its principal error term is the error term of the first method multiplied by $(-1)^p$.*

Proof. We replace h by $-h$ in (8.2), then $x \rightarrow x + h$ and rearrange the terms. This gives (using $d_{p+1}(x + h) = d_{p+1}(x) + \mathcal{O}(h)$)

$$y(x) + d_{p+1}(x)h^{p+1}(-1)^p + \mathcal{O}(h^{p+2})$$

$$= y(x + h) - h\Phi(x + h, y(x + h), -h).$$

Here we let B be the left-hand side of this identity, $A = y(x + h)$, and use (8.14). This leads to

$$y(x + h) = y(x) + d_{p+1}(x)h^{p+1}(-1)^p + h\Phi^*(x, y(x), h) + \mathcal{O}(h^{p+2}),$$

which expresses the statement of the theorem. \square

Theorem 8.6. *The adjoint method has exactly the same asymptotic expansion (8.10) as the original method, with h replaced by $-h$.*

Proof. We repeat the procedure which led to the proof of Theorem 8.1, with h negative. The first separated term corresponding to (8.9) will be

$$y(x) - y_{-h}(x) = e_p(x)(-h)^p + \mathcal{O}(h^{p+1}). \quad (8.9')$$

This is true because the solution of (8.8) with initial value $e_p(x_0) = 0$ has the same sign change as the inhomogeneity $d_{p+1}(x)$. This settles the first term. To continue, we prove that the transformation (8.4b) commutes with the adjunction operation, i.e., that

$$(\widehat{\Phi})^* = (\Phi^*)^\widehat{}. \quad (8.15)$$

In order to prove (8.15), we obtain from (8.4a) and the definition of $\widehat{\Phi}$

$$y_h(x+h) + e_p(x+h)h^p = y_h(x) + e_p(x)h^p + h\widehat{\Phi}(x, y_h(x) + e_p(x)h^p, h).$$

Here again, we substitute $h \rightarrow -h$ followed by $x \rightarrow x+h$. Finally, we apply (8.14) with $B = y_{-h}(x) + e_p(x)(-h)^p$ and $A = y_{-h}(x+h) + e_p(x+h)(-h)^p$ to obtain

$$\begin{aligned} y_{-h}(x+h) + e_p(x+h)(-h)^p \\ = y_{-h}(x) + e_p(x)(-h)^p + h(\widehat{\Phi})^*(x, y_{-h}(x) + e_p(x)(-h)^p, h). \end{aligned} \quad (8.16)$$

On the other hand, if we perform the transformation (see Theorem 8.5)

$$\widehat{y}_{-h}(x) = y_{-h}(x) + e_p(x)(-h)^p \quad (8.4')$$

and insert this into (8.13), we obtain (8.16) again, but this time with $(\Phi^*)^\widehat{}$ instead of $(\widehat{\Phi})^*$. This proves (8.15). \square

Symmetric Methods

Definition 8.7. A method is *symmetric* if $\Phi = \Phi^*$.

Example. The trapezoidal rule (7.5) and the implicit mid-point rule (7.4) are symmetric: the exchanges $y_1 \leftrightarrow y_0$, $h \leftrightarrow -h$ and $x_0 \leftrightarrow x_0 + h$ leave these methods invariant. The following two theorems (Wanner 1973) characterize symmetric IRK methods.

Theorem 8.8. *If*

$$a_{s+1-i, s+1-j} + a_{ij} = b_{s+1-j} = b_j, \quad i, j = 1, \dots, s, \quad (8.17)$$

then the corresponding Runge-Kutta method is symmetric. Moreover, if the b_i are nonzero and the c_i distinct and ordered as $c_1 < c_2 < \dots < c_s$, then condition (8.17) is also necessary for symmetry.

Proof. The sufficiency of (8.17) follows from Theorem 8.3. The condition $c_i = 1 - c_{s+1-i}$ can be verified by adding up (8.17) for $j = 1, \dots, s$.

Symmetry implies that the original method (with coefficients c_i, a_{ij}, b_j) and the adjoint method (c_i^*, a_{ij}^*, b_j^*) give identical numerical results. If we apply both methods to $y' = f(x)$ we obtain

$$\sum_{i=1}^s b_i f(c_i) = \sum_{i=1}^s b_i^* f(c_i^*)$$

for all $f(x)$. Our assumption on b_i and c_i thus yields

$$b_i^* = b_i, \quad c_i^* = c_i \quad \text{for all } i.$$

We next apply both methods to $y_1' = f(x)$, $y_2' = x^q y_1$ and obtain

$$\sum_{i,j=1}^s b_i c_i^q a_{ij} f(c_j) = \sum_{i,j=1}^s b_i^* c_i^{*q} a_{ij}^* f(c_j^*).$$

This implies $\sum_i b_i c_i^q a_{ij} = \sum_i b_i c_i^q a_{ij}^*$ for $q = 0, 1, \dots$ and hence also $a_{ij}^* = a_{ij}$ for all i, j . \square

Theorem 8.9. *A collocation method based on symmetrically distributed collocation points is symmetric.*

Proof. If $c_i = 1 - c_{s+1-i}$, the Lagrange polynomials satisfy $\ell_i(t) = \ell_{s+1-i}(1-t)$. Condition (8.17) is then an easy consequence of (7.19). \square

The following important property of symmetric methods, known intuitively for many years, now follows from the above results.

Theorem 8.10. *If in addition to the assumptions of Theorem 8.1 the underlying method is symmetric, then the asymptotic expansion (8.10) contains only even powers of h :*

$$y(x) - y_h(x) = e_{2q}(x)h^{2q} + e_{2q+2}(x)h^{2q+2} + \dots \quad (8.18)$$

with $e_{2j}(x_0) = 0$.

Proof. If $\Phi^* = \Phi$, we have $y_{-h}(x) = y_h(x)$ from (8.13) and the result follows from Theorem 8.6. \square

Exercises

1. Assume the one-step method (8.1) to be of order $p \geq 2$ and in addition to $\Phi(x, y, 0) = f(x, y)$ assume

$$\frac{\partial \Phi}{\partial h}(x, y, 0) = \frac{1}{2} \left(\frac{\partial f}{\partial x}(x, y) + \frac{\partial f}{\partial y}(x, y) \cdot f(x, y) \right). \quad (8.19)$$

Show that the principal local error term of the method $\hat{\Phi}$ defined in (8.5) is then given by

$$\hat{d}_{p+2}(x) = d_{p+2}(x) - \frac{1}{2} \frac{\partial f}{\partial y}(x, y(x)) d_{p+1}(x) - \frac{1}{2} d'_{p+1}(x).$$

Verify that (8.19) is satisfied for all RK-methods of order ≥ 2 .

2. Consider the second order method

0		
1		1
		1/2 1/2

applied to the problem $y' = y$, $y(0) = 1$. Show that

$$d_3(x) = \frac{1}{6} e^x, \quad d_4(x) = \frac{1}{24} e^x, \quad e_2(x) = \frac{1}{6} x e^x, \quad \hat{d}_4(x) = -\frac{1}{8} e^x.$$

3. Consider the second order method

0			
1/2	1/2		
1	0	1	
		1/4 1/2 1/4	

Show that for this method

$$d_3(x) = \frac{1}{24} \left(F(t_{32})(y(x)) - \frac{1}{2} F(t_{31})(y(x)) \right)$$

$$d_4(x) = \frac{1}{24} \left(F(t_{44})(y(x)) + \frac{1}{4} F(t_{43})(y(x)) - \frac{1}{4} F(t_{41})(y(x)) \right)$$

in the notation of Table 2.2. Show that this implies

$$\hat{d}_4(x) = 0 \quad \text{and} \quad e_3(x) = 0,$$

so that one step of Richardson extrapolation increases the order of the method by two. Find a connection between this method and the GBS-algorithm of Section II.9.

4. Discuss the symmetry of the IRK methods of Section II.7.

II.9 Extrapolation Methods

The following method of approximation may or may not be new, but as I believe it to be of practical importance . . .

(S.A. Corey 1906)

The h^2 -extrapolation was discovered by a hint from theory followed by arithmetical experiments, which gave pleasing results.

(L.F. Richardson 1927)

Extrapolation constitutes a powerful means . . .

(R. Bulirsch & J. Stoer 1966)

Extrapolation does not appear to be a particularly effective way . . . , our tests raise the question as to whether there is any point to pursuing it as a separate method.

(L.F. Shampine & L.S. Baca 1986)

Definition of the Method

Let $y' = f(x, y)$, $y(x_0) = y_0$ be a given differential system and $H > 0$ a basic step size. We choose a sequence of positive integers

$$n_1 < n_2 < n_3 < \dots \quad (9.1)$$

and define the corresponding step sizes $h_1 > h_2 > h_3 > \dots$ by $h_i = H/n_i$. We then choose a numerical method of order p and compute the numerical results of our initial value problem by performing n_i steps with step size h_i to obtain

$$y_{h_i}(x_0 + H) =: T_{i,1} \quad (9.2)$$

(the letter “ T ” stands historically for “trapezoidal rule”). We then eliminate as many terms as possible from the asymptotic expansion (8.10) by computing the interpolation polynomial

$$p(h) = \hat{y} - e_p h^p - e_{p+1} h^{p+1} - \dots - e_{p+k-2} h^{p+k-2} \quad (9.3)$$

such that

$$p(h_i) = T_{i,1} \quad i = j, j-1, \dots, j-k+1. \quad (9.4)$$

Finally we “*extrapolate to the limit*” $h \rightarrow 0$ and use

$$p(0) = \hat{y} =: T_{j,k}$$

as numerical result. Conditions (9.4) consist of k linear equations for the k unknowns $\hat{y}, e_p, \dots, e_{p+k-2}$.

Example. For $k = 2$, $n_1 = 1$, $n_2 = 2$ the above definition is identical to Richardson’s extrapolation discussed in Section II.4.

Theorem 9.1. *The value $T_{j,k}$ represents a numerical method of order $p+k-1$.*

Proof. We compare (9.4) and (9.3) with the asymptotic expansion (8.10) which we write in the form (with $N = p+k-1$)

$$T_{i,1} = y(x_0+H) - e_p(x_0+H)h_i^p - \dots - e_{p+k-2}(x_0+H)h_i^{p+k-2} - \Delta_i, \quad (9.4')$$

where

$$\Delta_i = e_{p+k-1}(x_0+H)h_i^{p+k-1} + E_{h_i}(x_0+H)h_i^{p+k} = \mathcal{O}(H^{p+k})$$

because $e_{p+k-1}(x_0) = 0$ and $h_i \leq H$. This is a linear system for the unknowns $y(x_0+H)$, $H^p e_p(x_0+H)$, \dots , $H^{p+k-2} e_{p+k-2}(x_0+H)$ with the Vandermonde-like matrix

$$A = \begin{pmatrix} 1 & \frac{1}{n_j^p} & \dots & \frac{1}{n_j^{p+k-2}} \\ \vdots & \vdots & & \vdots \\ 1 & \frac{1}{n_{j-k+1}^p} & \dots & \frac{1}{n_{j-k+1}^{p+k-2}} \end{pmatrix}.$$

It is the same as (9.4), just with the right-hand side perturbed by the $\mathcal{O}(H^{p+k})$ -terms Δ_i . The matrix A is invertible (see Exercise 6). Therefore by subtraction we obtain

$$|y(x_0+H) - \hat{y}| \leq \|A^{-1}\|_\infty \cdot \max |\Delta_i| = \mathcal{O}(H^{p+k}). \quad \square$$

Remark. The case $p=1$ (as well as $p=2$ with expansions in h^2) can also be treated by interpreting the difference $y(x_0+H) - \hat{y}$ as an interpolation error (see (9.21)).

A great advantage of the method is that it provides a complete table of numerical results

$$\begin{array}{ccccccc} T_{11} & & & & & & \\ T_{21} & T_{22} & & & & & \\ T_{31} & T_{32} & T_{33} & & & & \\ T_{41} & T_{42} & T_{43} & T_{44} & & & \\ \dots & \dots & \dots & \dots & \dots & & \end{array} \quad (9.5)$$

which form a sequence of embedded methods and allow easy estimates of the local error and strategies for variable order. Several step-number sequences are in use for (9.1):

The “Romberg sequence” (Romberg 1955):

$$1, 2, 4, 8, 16, 32, 64, 128, 256, 512, \dots \quad (9.6)$$

The “*Bulirsch sequence*” (see also Romberg 1955):

$$1, 2, 3, 4, 6, 8, 12, 16, 24, 32, \dots \quad (9.7)$$

alternating powers of 2 with 1.5 times 2^k . This sequence needs fewer function evaluations for higher orders than the previous one and became prominent through the success of the “Gragg-Bulirsch-Stoer algorithm” (Bulirsch & Stoer 1966).

The above sequences have the property that for integration problems $y' = f(x)$ many function values can be saved and re-used for smaller h_i . Further, $\liminf(n_{i+1}/n_i)$ remains bounded away from 1 (“Toeplitz condition”) which allows convergence proofs for $j = k \rightarrow \infty$ (Bauer, Rutishauser & Stiefel 1963). However, if we work with differential equations and with fixed or bounded order, the most economic sequence is the “*harmonic sequence*” (Deuffhard 1983)

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 10, \dots \quad (9.8)$$

The Aitken - Neville Algorithm

For the case $p = 1$, (9.3) and (9.4) become a classical interpolation problem and we can compute the values of $T_{j,k}$ economically by the use of classical methods. Since we need only the values of the interpolation polynomials at the point $h = 0$, the most economical algorithm is that of “Aitken - Neville” (Aitken 1932, Neville 1934, based on ideas of Jordan 1928) which leads to

$$T_{j,k+1} = T_{j,k} + \frac{T_{j,k} - T_{j-1,k}}{(n_j/n_{j-k}) - 1}. \quad (9.9)$$

If the basic method used is *symmetric*, we know that the underlying asymptotic expansion is in powers of h^2 (Theorem 8.9), and each extrapolation eliminates *two* powers of h . We may thus simply replace in (9.3) h by h^2 and for $p = 2$ (i.e., $q = 1$ in (8.18)) also use the Aitken - Neville algorithm with this modification. This leads to

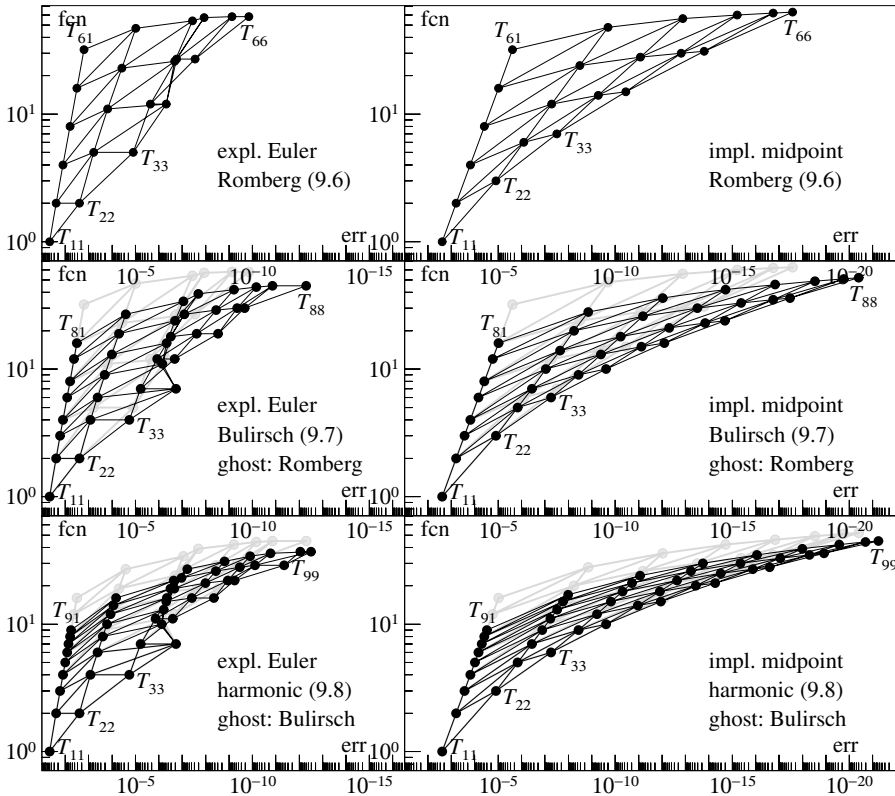
$$T_{j,k+1} = T_{j,k} + \frac{T_{j,k} - T_{j-1,k}}{(n_j/n_{j-k})^2 - 1} \quad (9.10)$$

instead of (9.9).

Numerical example. We solve the problem

$$y' = (-y \sin x + 2 \tan x)y, \quad y(\pi/6) = 2/\sqrt{3} \quad (9.11)$$

with true solution $y(x) = 1/\cos x$ and basic step size $H = 0.2$ by Euler’s method. Fig. 9.1 represents, for each of the entries $T_{j,k}$ of the extrapolation tableau, the *numerical work* $(1 + n_j - 1 + n_{j-1} - 1 + \dots + n_{j-k+1} - 1)$ compared to the *precision* $(|T_{j,k} - y(x_0 + H)|)$ in double logarithmic scale. The first picture is for the Romberg sequence (9.6), the second for the Bulirsch sequence (9.7), and the last

Fig. 9.1. h -extrap. expl. EulerFig. 9.2. h^2 -extrap. impl. midpoint

for the harmonic sequence (9.8). In pictures 2 and 3 the results of the foregoing graphics are repeated as a shaded “ghost” (. . . of Canterbury) in order to demonstrate how the results are better than those for the predecessor. Nobody is perfect, however. The “best” method in these comparisons, the harmonic sequence, suffers for high orders from a strong influence of rounding errors (see Exercise 5 below; the computations of Fig. 9.1, 9.2 and 9.4 have been made in quadruple precision).

The analogous results for the symmetric implicit mid-point rule (7.4) are presented in Fig. 9.2. Although implicit, this method is easy to implement for this particular example. We again use the same basic step size $H = 0.2$ as above and the same step-number sequences (9.6), (9.7), (9.8). Here, the “numerical work” $(n_j + n_{j-1} + \dots + n_{j-k+1})$ represents *implicit* stages and therefore can not be compared to the values of the explicit method. The precisions, however, show a drastic improvement.

Rational Extrapolation. Many authors in the sixties claimed that it is better to use rational functions instead of polynomials in (9.3). In this case the formula (9.9)

must be replaced by (Bulirsch & Stoer 1964)

$$T_{j,k+1} = T_{j,k} + \frac{T_{j,k} - T_{j-1,k}}{\left(\frac{n_j}{n_{j-k}}\right) \left(1 - \frac{T_{j,k} - T_{j-1,k}}{T_{j,k} - T_{j-1,k-1}}\right) - 1} \quad (9.12)$$

where

$$T_{j,0} = 0.$$

For systems of differential equations the division of vectors is to be understood componentwise.

Later numerical experiments (Deuffhard 1983) showed that rational extrapolation is nearly never more advantageous than polynomial extrapolation.

The Gragg or GBS Method

Since it is fully explicit GRAGG's algorithm is so ideally suited as a basis for RICHARDSON extrapolation that no other symmetric two-step algorithm can compete with it. (H.J. Stetter 1970)

Here we can not do better than quote from Stetter (1970): "Expansions in powers of h^2 are extremely important for an efficient application of Richardson extrapolation. Therefore it was a great achievement when Gragg proved in 1963 that the quantity $S_h(x)$ produced by the algorithm ($x = x_0 + 2nh$, $x_i = x_0 + ih$)

$$y_1 = y_0 + hf(x_0, y_0) \quad (9.13a)$$

$$y_{i+1} = y_{i-1} + 2hf(x_i, y_i) \quad i = 1, 2, \dots, 2n \quad (9.13b)$$

$$S_h(x) = \frac{1}{4} (y_{2n-1} + 2y_{2n} + y_{2n+1}) \quad (9.13c)$$

possesses an asymptotic expansion in even powers of h and has satisfactory stability properties. This led to the construction of the very powerful G(ragg)-B(ulirsch)-S(toer)-extrapolation algorithm . . .".

Gragg's *proof* of this property was very long and complicated and it was again "a great achievement" that Stetter had the elegant idea of interpreting (9.13b) as a *one-step* algorithm by rewriting (9.13) in terms of odd and even indices: for this purpose we define

$$\begin{aligned} h^* &= 2h, & x_k^* &= x_0 + kh^*, & u_0 &= v_0 = y_0, \\ u_k &= y_{2k}, & v_k &= y_{2k+1} - hf(x_{2k}, y_{2k}) = \frac{1}{2} (y_{2k+1} + y_{2k-1}). \end{aligned} \quad (9.14)$$

Then the method (9.13) can be rewritten as (see Fig. 9.3)

$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} u_k \\ v_k \end{pmatrix} + h^* \begin{pmatrix} f\left(x_k^* + \frac{h^*}{2}, v_k + \frac{h^*}{2} f(x_k^*, u_k)\right) \\ \frac{1}{2} \left(f(x_k^* + h^*, u_{k+1}) + f(x_k^*, u_k)\right) \end{pmatrix}. \quad (9.15)$$

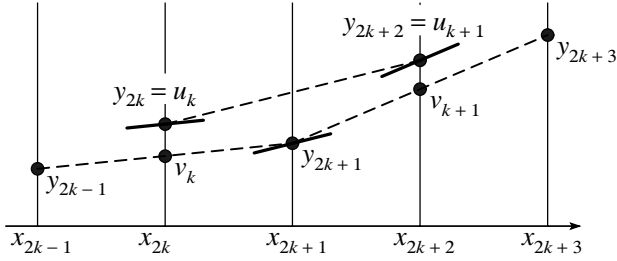


Fig. 9.3. Symmetry of the Gragg method

This method, which maps the pair (u_k, v_k) to (u_{k+1}, v_{k+1}) , can be seen from Fig. 9.3 to be *symmetric*. The symmetry can also be checked analytically (see Definition 8.7) by exchanging $u_{k+1} \leftrightarrow u_k$, $v_{k+1} \leftrightarrow v_k$, $h^* \leftrightarrow -h^*$, $x_k^* \leftrightarrow x_k^* + h^*$. A trivial calculation then shows that this leaves formula (9.15) invariant. Method (9.15) is consistent with the differential equation (let $h^* \rightarrow 0$ in the increment function)

$$\begin{aligned} u' &= f(x, v) & u(x_0) &= y_0 \\ v' &= f(x, u) & v(x_0) &= y_0, \end{aligned} \quad (9.16)$$

whose exact solution is simply $u(x) = v(x) = y(x)$. Therefore, we have from Theorem 8.10 that

$$y(x) - u_{h^*}(x) = \sum_{j=1}^{\ell} a_{2j}(x)(h^*)^{2j} + (h^*)^{2\ell+2} A(x, h^*) \quad (9.17a)$$

$$y(x) - v_{h^*}(x) = \sum_{j=1}^{\ell} b_{2j}(x)(h^*)^{2j} + (h^*)^{2\ell+2} B(x, h^*) \quad (9.17b)$$

and $a_{2j}(x_0) = b_{2j}(x_0) = 0$. We see from (9.14) and (9.17a) that $y_h(x)$ possesses an expansion in even powers of h , provided that the number of steps is even; i.e., for $x = x_0 + 2nh$,

$$y(x) - y_h(x) = \sum_{j=1}^{\ell} \hat{a}_{2j}(x) h^{2j} + h^{2\ell+2} \hat{A}(x, h) \quad (9.18)$$

where $\hat{a}_{2j}(x) = 2^{2j} a_{2j}(x)$ and $\hat{A}(x, h) = 2^{2\ell+2} A(x, 2h)$.

The so-called *smoothing step*, i.e., formula

$$S_h(x_0 + 2nh) = \frac{1}{4} (y_{2n-1} + 2y_{2n} + y_{2n+1}) = \frac{1}{2} (u_n + v_n)$$

(see (9.13c) and (9.14)) had its historical origin in the “weak stability” of the explicit midpoint rule (9.13b) (see also Fig. III.9.2). However, since the method is anyway followed by extrapolation, this step is not of great importance (Shampine & Baca 1983). It is a little more costly and increases the “stability domain” by

approximately the same amount (see Fig. IV.2.3 of Vol. II). Further, it has the advantage of evaluating the function f at the end of the basic step.

Theorem 9.2. *Let $f(x, y) \in \mathcal{C}^{2\ell+2}$, then the numerical solution defined in (9.13) possesses for $x = x_0 + 2nh$ an asymptotic expansion of the form*

$$y(x) - S_h(x) = \sum_{j=1}^{\ell} e_{2j}(x)h^{2j} + h^{2\ell+2}C(x, h) \quad (9.19)$$

with $e_{2j}(x_0) = 0$ and $C(x, h)$ bounded for $x_0 \leq x \leq \bar{x}$ and $0 \leq h \leq h_0$.

Proof. By adding (9.17a) and (9.17b) and using $h^* = 2h$ we obtain (9.19) with $e_{2j}(x) = (a_{2j}(x) + b_{2j}(x))2^{2j-1}$. \square

This method can thus be used for Richardson extrapolation in the same way as symmetric methods above: we choose a step-number sequence, with the condition that the n_j are even, i.e.,

$$2, 4, 8, 16, 32, 64, 128, 256, \dots \quad (9.6')$$

$$2, 4, 6, 8, 12, 16, 24, 32, 48, \dots \quad (9.7')$$

$$2, 4, 6, 8, 10, 12, 14, 16, 18, \dots \quad (9.8')$$

set

$$T_{i,1} := S_{h_i}(x_0 + H)$$

and compute the extrapolated expressions $T_{i,j}$, based on the h^2 -expansion, by the Aitken-Neville formula (9.10).

Numerical example. Fig. 9.4 represents the numerical results of this algorithm applied to Example (9.11) with step size $H = 0.2$. The step size sequences are Romberg (9.6') (above), Bulirsch (9.7') (middle), and harmonic (9.8') (below). The algorithm *with* smoothing step (numerical work $= 1 + n_j + n_{j-1} + \dots + n_{j-k+1}$) is represented *left*, the results *without* smoothing step (numerical work $= 1 + n_j - 1 + n_{j-1} - 1 + \dots + n_{j-k+1} - 1$) are on the *right*.

The results are nearly identical to those for the implicit midpoint rule (Fig. 9.2), but much more valuable, since here the method is explicit. In the pictures on the left the values for extrapolated Euler (from Fig. 9.1) are repeated as a “ghost” and demonstrate clearly the importance of the h^2 -expansion, especially in the diagonal T_{kk} for large values of k . The ghost in the pictures on the right are the values *with* smoothing from the left; the differences are seen to be tiny.

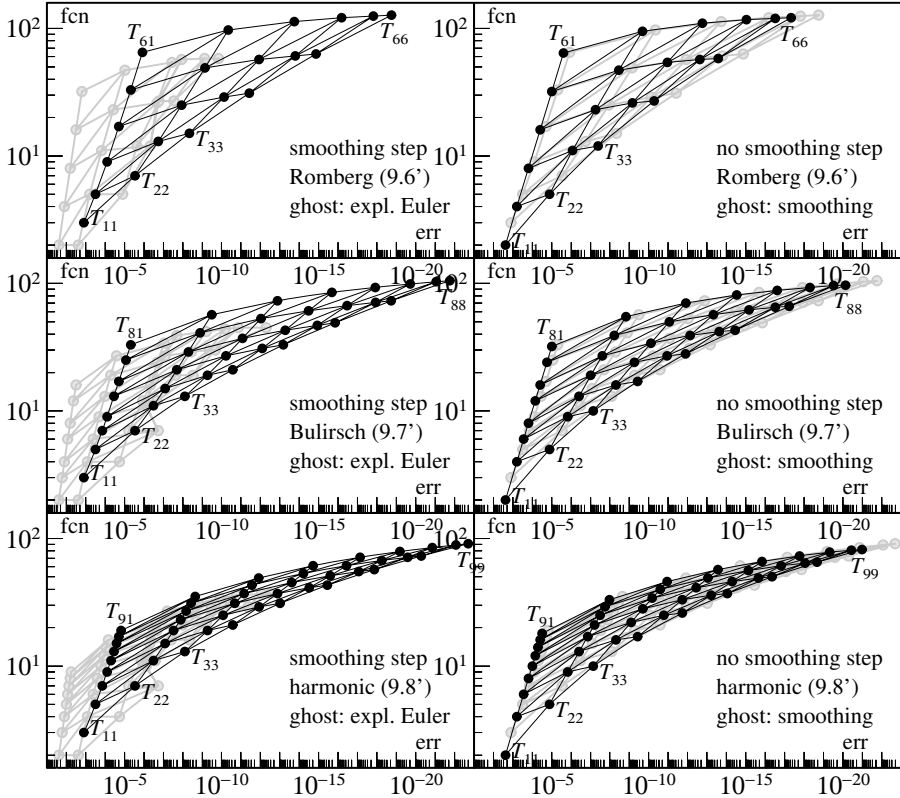


Fig. 9.4. Precision of h^2 -extrapolated Gragg method for Example (9.11)

Asymptotic Expansion for Odd Indices

For completeness, we still want to derive the existence of an h^2 expansion for y_{2k+1} from (9.17b), although this is of no practical importance for the numerical algorithm described above.

Theorem 9.3 (Gragg 1964). *For $x = x_0 + (2k+1)h$ we have*

$$y(x) - y_h(x) = \sum_{j=1}^{\ell} \hat{b}_{2j}(x) h^{2j} + h^{2\ell+2} \hat{B}(x, h) \quad (9.20)$$

where the coefficients $\hat{b}_{2j}(x)$ are in general different from those for even indices and $\hat{b}_{2j}(x_0) \neq 0$.

Proof. y_{2k+1} can be computed (see Fig. 9.3) either from v_k by a forward step or from v_{k+1} by a backward step. For the sake of symmetry, we take the mean of

both expressions and write

$$y_{2k+1} = \frac{1}{2}(v_k + v_{k+1}) + \frac{h}{2}(f(x_k^*, u_k) - f(x_{k+1}^*, u_{k+1})).$$

We now subtract the exact solution and obtain

$$\begin{aligned} 2(y_h(x) - y(x)) &= v_{2h}(x-h) - y(x-h) \\ &\quad + v_{2h}(x+h) - y(x+h) + y(x-h) - 2y(x) + y(x+h) \\ &\quad + h\left(f(x-h, u_{2h}(x-h)) - f(x+h, u_{2h}(x+h))\right). \end{aligned}$$

Due to the symmetry of $u_{2h}(x)$ ($u_{2h}(\xi) = u_{-2h}(\xi)$) and of $v_{2h}(x)$ the whole expression becomes symmetric in h . Thus the asymptotic expansion for y_{2k+1} contains no odd powers of h . \square

Both expressions, for even and for odd indices, can still be combined into a single formula (see Exercise 2).

Existence of Explicit RK Methods of Arbitrary Order

Each of the expressions $T_{j,k}$ clearly represents an explicit RK-method (see Exercise 1). If we apply the well-known error formula for polynomial interpolation (see e.g., Abramowitz & Stegun 1964, formula 25.2.27) to (9.19), we obtain

$$y(x_0 + H) - T_{j,k} = \frac{(-1)^k}{n_j^2 \cdots n_{j-k+1}^2} e_{2k}(x_0 + H) H^{2k} + \mathcal{O}(H^{2k+2}). \quad (9.21)$$

Since $e_k(x_0) = 0$, we have

$$y(x_0 + H) - T_{j,k} = \frac{(-1)^k}{n_j^2 \cdots n_{j-k+1}^2} e'_{2k}(x_0) H^{2k+1} + \mathcal{O}(H^{2k+2}). \quad (9.22)$$

This shows that $T_{j,k}$ represents an explicit Runge-Kutta method of order $2k$. As an application of this result we have:

Theorem 9.4 (Gragg 1964). *For p even, there exists an explicit RK-method of order p with $s = p^2/4 + 1$ stages.*

Proof. This result is obtained by counting the number of necessary function evaluations of the GBS-algorithm using the harmonic sequence and without the final smoothing step. \square

Remark. The extrapolated Euler method leads to explicit Runge-Kutta methods with $s = p(p-1)/2 + 1$ stages. This shows once again the importance of the h^2 expansion.

Order and Step Size Control

Extrapolation methods have the advantage that in addition to the step size also the order (i.e., number of columns) can be changed at each step. Because of this double freedom, the *practical implementation* in an optimal way is more complicated than for fixed-order RK-methods. The first codes were developed by Bulirsch & Stoer (1966) and their students. Very successful extrapolation codes due to P. Deuffhard and his collaborators are described in Deuffhard (code DIFEX1, 1983).

The choice of the *step size* can be performed in exactly the same way as for fixed-order embedded methods (see Section II.4). If the first k lines of the extrapolation tableau are computed, we have T_{kk} as the highest-order approximation (of order $2k$ by (9.22)) and in addition $T_{k,k-1}$ of order $2k-2$. It is therefore natural to use the expression

$$err_k = \|T_{k,k-1} - T_{k,k}\| \quad (9.23)$$

for step size control. The norm is the same as in (4.11). As in (4.12) we get for the optimal step size the formula

$$H_k = H \cdot 0.94 \cdot (0.65/err_k)^{1/(2k-1)} \quad (9.24)$$

where this time we have chosen a safety factor depending partly on the order.

For the choice of an *optimal order* we need a measure of work, which allows us to compare different methods. The work for computing T_{kk} can be measured by the number A_k of function evaluations. For the GBS-algorithm it is given recursively by

$$\begin{aligned} A_1 &= n_1 + 1 \\ A_k &= A_{k-1} + n_k. \end{aligned} \quad (9.25)$$

However, a large number of function evaluations can be compensated by a large step size H_k , given by (9.24). We therefore consider

$$W_k = \frac{A_k}{H_k}, \quad (9.26)$$

the *work per unit step*, as a measure of work. The idea is now to choose the order (i.e., the index k) in such a way that W_k is minimized.

Let us describe the *combined order and step size control* in some more detail. We assume that at some point of integration the step size H and the index k ($k > 2$) are proposed. The step is then realized in the following way: we first compute $k-1$ lines of the extrapolation tableau and also the values H_{k-2} , W_{k-2} , err_{k-1} , H_{k-1} , W_{k-1} .

a) *Convergence in line $k-1$.* If $err_{k-1} \leq 1$, we accept $T_{k-1,k-1}$ as numerical solution and continue the integration with the new proposed quantities

$$k_{\text{new}} = \begin{cases} k & \text{if } W_{k-1} < 0.9 \cdot W_{k-2} \\ k-1 & \text{else} \end{cases} \quad (9.27)$$

$$H_{\text{new}} = \begin{cases} H_{k_{\text{new}}} & \text{if } k_{\text{new}} \leq k-1 \\ H_{k-1}(A_k/A_{k-1}) & \text{if } k_{\text{new}} = k. \end{cases}$$

In (9.27), the only non-trivial formula is the choice of the step size H_{new} in the case of an order-increase $k_{\text{new}} = k$. In this case we want to avoid the computation of err_k , so that H_k and W_k are unknown. However, since our k is assumed to be close to the optimal value, we have $W_k \approx W_{k-1}$ which leads to the proposed step size increase.

b) *Convergence monitor.* If $err_{k-1} > 1$, we first decide whether we may expect convergence at least in line $k+1$. It follows from (9.22) that, asymptotically,

$$\|T_{k,k-2} - T_{k,k-1}\| \approx \left(\frac{n_2}{n_k}\right)^2 err_{k-1} \quad (9.28)$$

with err_{k-1} given by (9.23). Unfortunately, err_k cannot be compared with (9.28), since different factors (depending on the differential equation to be solved) are involved in the asymptotic formula (cf. (9.22)). If we nevertheless assume that err_k is $(n_2/n_1)^2$ times smaller than (9.28) we obtain $err_k \approx (n_1/n_k)^2 err_{k-1}$. We therefore already reject the step at this point, if

$$err_{k-1} > \left(\frac{n_{k+1}n_k}{n_1n_1}\right)^2 \quad (9.29)$$

and restart with $k_{\text{new}} \leq k-1$ and H_{new} according to (9.27). If the contrary of (9.29) holds, we compute the next line of the extrapolation tableau, i.e., $T_{k,k}$, err_k , H_k and W_k .

c) *Convergence in line k .* If $err_k \leq 1$, we accept T_{kk} as numerical solution and continue the integration with the new proposed values

$$k_{\text{new}} = \begin{cases} k-1 & \text{if } W_{k-1} < 0.9 \cdot W_k \\ k+1 & \text{if } W_k < 0.9 \cdot W_{k-1} \\ k & \text{in all other cases} \end{cases} \quad (9.30)$$

$$H_{\text{new}} = \begin{cases} H_{k_{\text{new}}} & \text{if } k_{\text{new}} \leq k \\ H_k(A_{k+1}/A_k) & \text{if } k_{\text{new}} = k+1. \end{cases}$$

d) *Second convergence monitor.* If $err_k > 1$, we check, as in (b), the relation

$$err_k > \left(\frac{n_{k+1}}{n_1}\right)^2. \quad (9.31)$$

If (9.31) is satisfied, the step is rejected and we restart with $k_{\text{new}} \leq k$ and H_{new} of (9.30). Otherwise we continue.

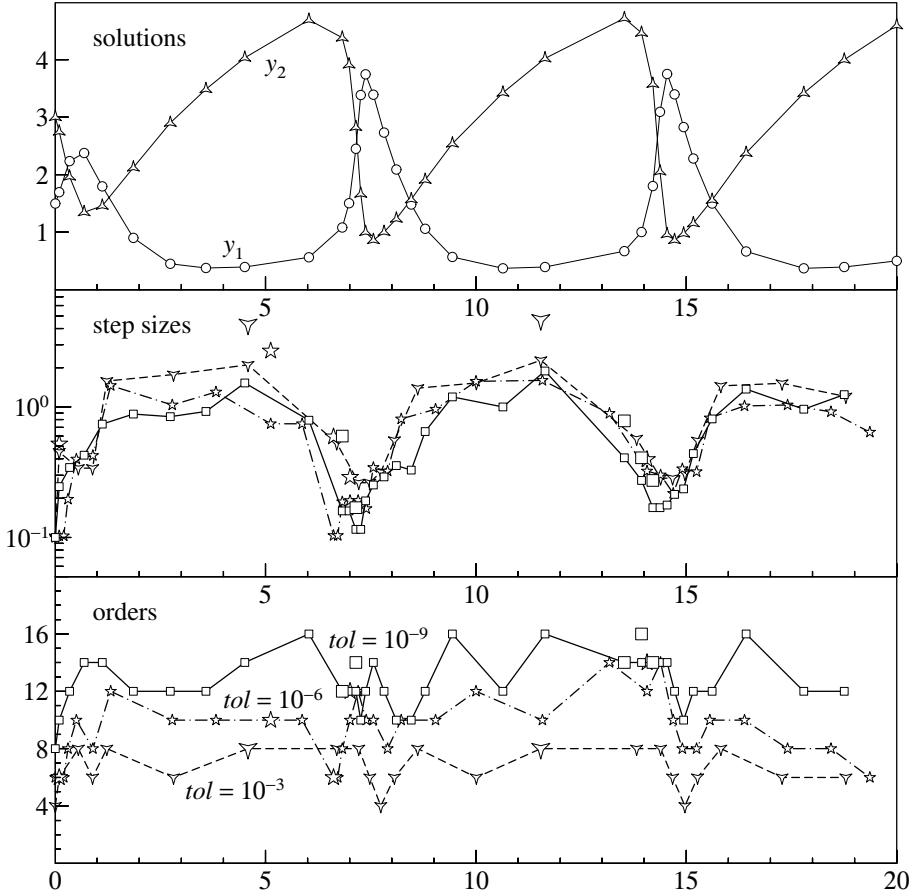


Fig. 9.5. Solution, step size and order variation obtained by ODEX

e) *Hope for convergence in line $k+1$.* We compute err_{k+1} , H_{k+1} and W_{k+1} . If $err_{k+1} \leq 1$, we accept $T_{k+1,k+1}$ as numerical solution and continue the integration with the new proposed order

$$\begin{aligned}
 k_{\text{new}} &:= k \\
 \text{if } (W_{k-1} < 0.9 \cdot W_k) & \quad k_{\text{new}} := k-1 \\
 \text{if } (W_{k+1} < 0.9 \cdot W_{k_{\text{new}}}) & \quad k_{\text{new}} := k+1.
 \end{aligned} \tag{9.32}$$

If $err_{k+1} > 1$ the step is rejected and we restart with $k_{\text{new}} \leq k$ and H_{new} of (9.24).

The following slight modifications of the above algorithm are recommended:

i) Storage considerations lead to a limitation of the number of columns of the extrapolation tableau, say by k_{max} (e.g., $k_{\text{max}} = 9$). For the proposed index k_{new} we require $2 \leq k_{\text{new}} \leq k_{\text{max}} - 1$. This allows us to activate (e) at each step.

ii) After a step-rejection the step size and the order may not be increased.

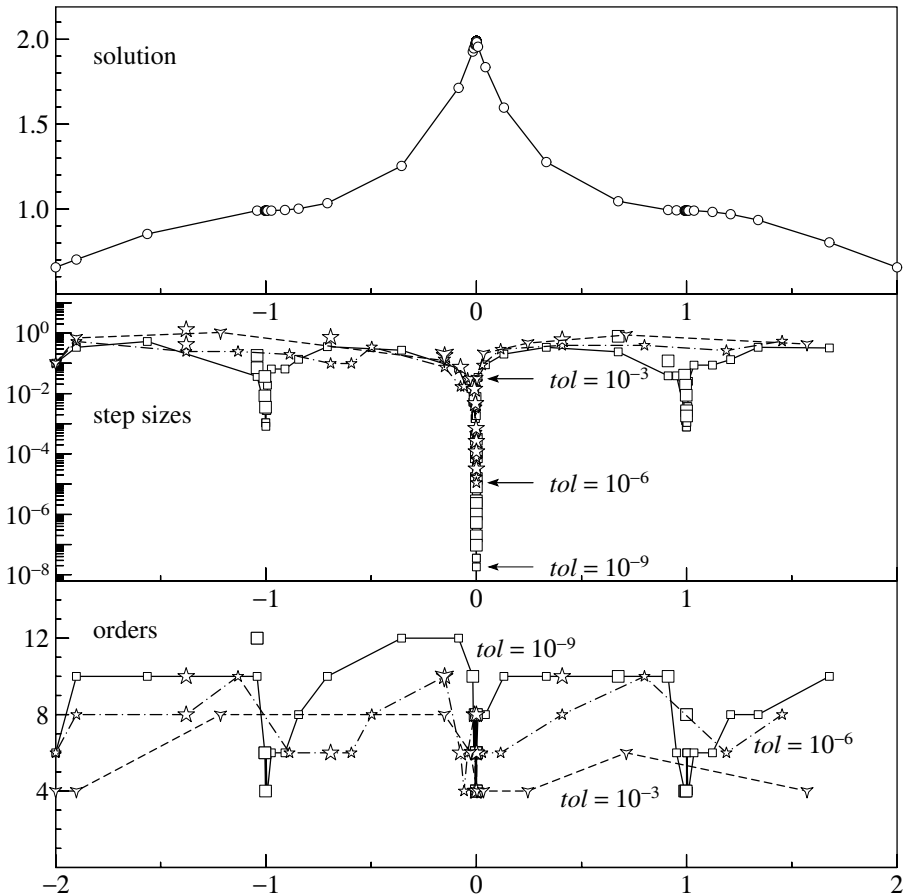


Fig. 9.6. Solution, step size and order variation obtained by ODEX at the discontinuous example (9.33)

Numerical study of the combined step size and order control. We show in the following examples how the step size and the order vary for the above algorithm. For this purpose we have written the FORTRAN-subroutine ODEX (see Appendix).

As a first example we again take the *Brusselator* (cf. Section II.4). As in Fig. 4.1, the first picture of Fig. 9.5 shows the two components of the solution (obtained with $Atol = Rtol = 10^{-9}$). In the remaining two pictures we have plotted the step sizes and orders for the three tolerances 10^{-3} (broken line), 10^{-6} (dashes and dots) and 10^{-9} (solid line). One can easily observe that the extrapolation code automatically chooses a suitable order (depending essentially on Tol). Step-rejections are indicated by larger symbols.

We next study the behaviour of the order control near discontinuities. In the example

$$y' = -\text{sign}(x) |1 - |x|| y^2, \quad y(-2) = 2/3, \quad -2 \leq x \leq 2 \quad (9.33)$$

we have a discontinuity in the first derivative of $y(x)$ at $x = 0$ and two discontinuities in the second derivative (at $x = \pm 1$). The numerical results are shown in Fig. 9.6 for three tolerances. In all cases the error at the endpoint is about $10 \cdot \text{Tot}$. The discontinuities at $x = \pm 1$ are not recognized in the computations with $\text{Tot} = 10^{-3}$ and $\text{Tot} = 10^{-6}$. Whenever a discontinuity is detected, the order drops to 4 (lowest possible) in its neighbourhood, so that these points are passed rather efficiently.

Dense Output for the GBS Method

Extrapolation methods are methods best suited for high precision which typically take very large (basic) step sizes during integration. The reasons for the need of a dense output formula (discussed in Section II.6) are therefore particularly important here. First attempts to provide extrapolation methods with a dense output are due to Lindberg (1972) for the implicit trapezoidal rule, and to Shampine, Baca & Bauer (1983) who constructed a 3rd order dense output for the GBS method. We present here the approach of Hairer & Ostermann (1990) (see also Simonsen 1990).

It turned out that the existence of high order dense output is only possible if the step number sequence satisfies some restrictions such as

$$n_{j+1} - n_j = 0 \pmod{4} \quad \text{for } j = 1, 2, 3, \dots \quad (9.34)$$

which, for example, is fulfilled by the sequence

$$\{2, 6, 10, 14, 18, 22, 26, 30, 34, \dots\}. \quad (9.35)$$

The idea is, once again, to do Hermite interpolation. To begin with, high order approximations are as usual at our disposal for the values y_0, y'_0, y_1, y'_1 by using $y_0, f(x_0, y_0), T_{kk}, f(x_0 + H, T_{kk})$, where T_{kk} is supposed to be the highest order approximation computed and used for continuation of the solution.

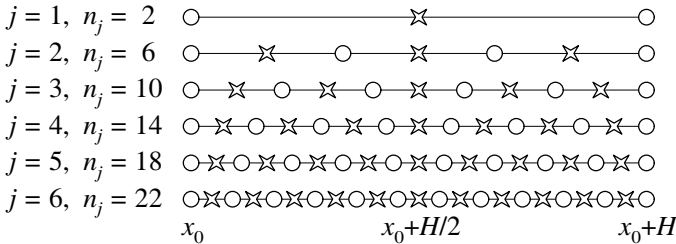


Fig. 9.7. Evaluation points for a GBS step

For more inspiration, we represent in Fig. 9.7 the steps taken by Gragg's midpoint rule for the step number sequence (9.35). The symbols \circ and \times indicate that the even steps and the odd steps possess a *different* asymptotic expansion (see Theorem 9.3) and must not be blended. We see that, owing to condition (9.34), the midpoint values $y_{n_j/2}^{(j)}$, obtained during the computation of T_{j1} , all have the same parity and can therefore also be extrapolated to yield an approximation for $y(x_0 + H/2)$ of order $2k - 1$ (remember that in Theorem 9.3, $\widehat{b}_{2j}(x_0) \neq 0$).

We next insert (9.20) for $x = x_0 + H/2$ into $f(x, y)$

$$f_{n_j/2}^{(j)} := f(x, y_{n_j/2}^{(j)}) = f\left(x, y(x) - h_j^2 \widehat{b}_2(x) - h_j^4 \widehat{b}_4(x) \dots\right)$$

and develop in powers of h_j to obtain

$$y'(x) - f_{n_j/2}^{(j)} = h_j^2 a_{2,1}(x) + h_j^4 a_{4,1}(x) + \dots \quad (9.36)$$

This shows that the f -values at the midpoint $x_0 + H/2$ (for $j = 1, 2, \dots, k$) possess an asymptotic expansion and can be extrapolated $k - 1$ times to yield an approximation to $y'(x_0 + H/2)$ of order $2k - 1$.

But this is not enough. We now consider, similar to an idea which goes back to the papers of Deuflhard & Nowak (1987) and Lubich (1989), the central differences $\delta f_i = f_{i+1} - f_{i-1}$ at the midpoint which, by Fig. 9.7, are available for $j = 1, 2, \dots, k$ and are based on even parity. By using (9.18) and by developing into powers of h_j we obtain

$$\begin{aligned} \frac{\delta f_{n_j/2}^{(j)}}{2h_j} &= \frac{f(x + h_j, y_{n_j/2+1}^{(j)}) - f(x - h_j, y_{n_j/2-1}^{(j)})}{2h_j} \\ &= \left(f(x + h_j, y(x + h_j) - h_j^2 \widehat{a}_2(x + h_j) - h_j^4 \widehat{a}_4(x + h_j) - \dots) - \right. \\ &\quad \left. f(x - h_j, y(x - h_j) - h_j^2 \widehat{a}_2(x - h_j) - h_j^4 \widehat{a}_4(x - h_j) - \dots) \right) / 2h_j \\ &= \frac{y'(x + h_j) - y'(x - h_j)}{2h_j} - h_j^2 c_2(x) - h_j^4 c_4(x) - \dots \end{aligned}$$

Finally we insert the Taylor series for $y'(x + h)$ and $y'(x - h)$ to obtain an expansion

$$y''(x) - \frac{\delta f_{n_j/2}^{(j)}}{2h_j} = h_j^2 a_{2,2}(x) + h_j^4 a_{4,2}(x) + \dots \quad (9.38)$$

Therefore, $k - 1$ extrapolations of the expressions (9.37) yield an approximation to $y''(x_0 + H/2)$ of order $2k - 1$.

In order to get approximations to the third and fourth derivatives of the solution at $x_0 + H/2$, we use the second and third central differences of $f_i^{(j)}$ which exist for $j \geq 2$ (Fig. 9.7). These can be extrapolated $k - 2$ times to give approximations of order $2k - 3$.

The continuation of this process yields the following algorithm:

Step 1. For each $j \in \{1, \dots, k\}$, compute approximations to the derivatives of $y(x)$ at $x_0 + H/2$ by:

$$d_j^{(0)} = y_{n_j/2}^{(j)}, \quad d_j^{(\kappa)} = \frac{\delta^{\kappa-1} f_{n_j/2}^{(j)}}{(2h_j)^{\kappa-1}} \quad \text{for } \kappa = 1, \dots, 2j. \quad (9.39)$$

Step 2. Extrapolate $d_j^{(0)}$ $(k-1)$ times and $d_j^{(2\ell-1)}$, $d_j^{(2\ell)}$ $(k-\ell)$ times to obtain improved approximations $d^{(\kappa)}$ to $y^{(\kappa)}(x_0 + H/2)$.

Step 3. For given μ ($-1 \leq \mu \leq 2k$) define the polynomial $P_\mu(\theta)$ of degree $\mu + 4$ by

$$\begin{aligned} P_\mu(0) &= y_0, & P'_\mu(0) &= Hf(x_0, y_0), \\ P_\mu(1) &= T_{kk}, & P'_\mu(1) &= Hf(x_0 + H, T_{kk}) \\ P_\mu^{(\kappa)}(1/2) &= H^\kappa d^{(\kappa)} & \text{for } \kappa &= 0, \dots, \mu. \end{aligned} \quad (9.40)$$

This computation of $P_\mu(\theta)$ does not need any further function evaluation since $f(x_0 + H, T_{kk})$ has to be computed anyway for the next step. Further, $P_\mu(\theta)$ gives a global \mathcal{C}^1 approximation to the solution.

Theorem 9.5 (Hairer & Ostermann 1990). *If the step number sequence satisfies (9.34), then the error of the dense output polynomial $P_\mu(\theta)$ satisfies*

$$y(x_0 + \theta H) - P_\mu(\theta) = \begin{cases} \mathcal{O}(H^{2k+1}) & \text{if } n_1 = 4 \text{ and } \mu \geq 2k-4 \\ \mathcal{O}(H^{2k}) & \text{if } n_1 = 2 \text{ and } \mu \geq 2k-5. \end{cases} \quad (9.40)$$

Proof. Since $P_\mu(\theta)$ is a polynomial of degree $\mu + 4$ the error due to interpolation is of size $\mathcal{O}(H^{\mu+5})$. This explains the restriction on μ in (9.40). As explained above, the function value and derivative data used for Hermite interpolation have the required precision

$$H^\kappa y^{(\kappa)}(x_0 + H/2) - H^\kappa d^{(\kappa)} = \begin{cases} \mathcal{O}(H^{2k}) & \text{if } \kappa = 0, \\ \mathcal{O}(H^{2k+1}) & \text{if } \kappa \text{ is odd,} \\ \mathcal{O}(H^{2k+2}) & \text{if } \kappa \geq 2 \text{ is even.} \end{cases}$$

In the case $n_1 = 4$ the parity of the central point $x_0 + H/2$ is *even* (in contrary to Fig. 9.7), we therefore apply (9.18) and gain one order because then the functions $a_{i,0}(x)$ vanish at x_0 . \square

Control of the Interpolation Error

At one time . . . every young mathematician was familiar with $\operatorname{sn} u$, $\operatorname{cn} u$, and $\operatorname{dn} u$, and algebraic identities between these functions figured in every examination.

(E.H. Neville, *Jacobian Elliptic Functions*, 1944)

Numerical example. We apply the above dense output formula with $\mu = 2k - 3$ (as is standard in ODEX) to the differential equations of the Jacobian elliptic functions sn , cn , dn (see Abramowitz & Stegun 1964, 16.16):

$$\begin{aligned} y_1' &= y_2 y_3 & y_1(0) &= 0 \\ y_2' &= -y_1 y_3 & y_2(0) &= 1 \\ y_3' &= -0.51 \cdot y_1 y_2 & y_3(0) &= 1 \end{aligned} \quad (9.41)$$

with integration interval $0 \leq x \leq 10$ and error tolerance $Atol = Rtol = 10^{-9}$. The error for the three components of the obtained continuous solution is displayed in Fig. 9.8 (upper picture; the ghosts are the solution curves) and gives a quite disappointing impression when compared with the precision at the grid points. We shall now see that these horrible bumps are nothing else than interpolation errors.

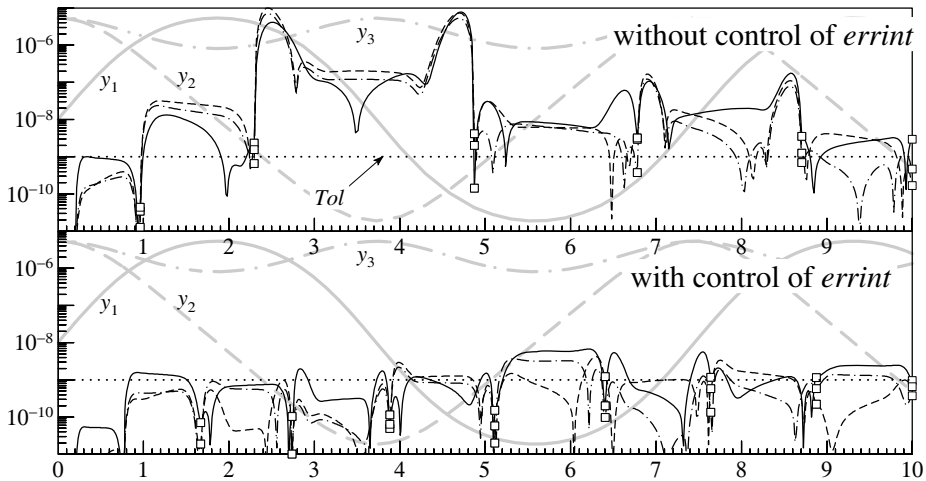


Fig. 9.8. Error of dense output without/with interpolation control

Assume that in the definition of $P_\mu(\theta)$ the basic function and derivative values are replaced by the exact values $y(x_0 + H)$, $y'(x_0 + H)$, and $y^{(\kappa)}(x_0 + H/2)$. Then the error of $P_\mu(\theta)$ is given by

$$\theta^2(1-\theta)^2 \left(\theta - \frac{1}{2} \right)^{\mu+1} \frac{y^{(\mu+5)}(\xi)}{(\mu+5)!} H^{\mu+5} \quad (9.42)$$

where $\xi \in (x_0, x_0 + H)$ (possibly different for each component). The function $\theta^2(1 - \theta)^2(\theta - 1/2)^{\mu+1}$ has its maximum at

$$\theta_{\mu+1} = \frac{1}{2} \pm \frac{1}{2} \sqrt{\frac{\mu+1}{\mu+5}} \quad (9.43)$$

which, for large μ , are close to the ends of the integration intervals and indicate precisely the locations of the large bumps in Fig. 9.8. This demonstrates the need for a code which not only controls the error at the grid points, but also takes care of the interpolation error. To this end we denote by a_μ the coefficient of $\theta^{\mu+4}$ in the polynomial $P_\mu(\theta)$ and consider (Hairer & Ostermann 1992)

$$P_\mu(\theta) - P_{\mu-1}(\theta) = \theta^2(1 - \theta)^2 \left(\theta - \frac{1}{2}\right)^\mu a_\mu \quad (9.44)$$

as an approximation for the interpolation error for $P_{\mu-1}(\theta)$ and use

$$errint = \|P_\mu(\theta_\mu) - P_{\mu-1}(\theta_\mu)\| \quad (9.45)$$

as error estimator (the norm is again that of (4.11)). Then, if $errint > 10$ the step is rejected and recomputed with

$$H_{\text{int}} = H(1/errint)^{1/(\mu+4)}$$

because $errint = \mathcal{O}(H^{\mu+4})$. Otherwise the subsequent step is computed subject to the restriction $H \leq H_{\text{int}}$.

This modified step size strategy makes the code, together with its dense output, more robust. The corresponding numerical results for the problem (9.41) are presented in the lower graph of Fig. 9.8.

Exercises

1. Show that the extrapolated Euler methods $T_{3,1}, T_{3,2}, T_{3,3}$ (with step-number sequence (9.8)) are equivalent to the Runge-Kutta methods of Table 9.1. Compute also the Runge-Kutta schemes corresponding to the first elements of the GBS algorithm.

Table 9.1. Extrapolation methods as Runge-Kutta methods

<table> <tr><td>0</td><td></td></tr> <tr><td>1/3</td><td>1/3</td></tr> <tr><td>2/3</td><td>1/3 1/3</td></tr> <tr><td colspan="2">1/3 1/3 1/3</td></tr> </table>	0		1/3	1/3	2/3	1/3 1/3	1/3 1/3 1/3		<table> <tr><td>0</td><td></td></tr> <tr><td>1/2</td><td>1/2</td></tr> <tr><td>1/3</td><td>1/3 0</td></tr> <tr><td>2/3</td><td>1/3 0 1/3</td></tr> <tr><td colspan="2">0 -1 1 1</td></tr> </table>	0		1/2	1/2	1/3	1/3 0	2/3	1/3 0 1/3	0 -1 1 1		<table> <tr><td>0</td><td></td></tr> <tr><td>1/2</td><td>1/2</td></tr> <tr><td>1/3</td><td>1/3 0</td></tr> <tr><td>2/3</td><td>1/3 0 1/3</td></tr> <tr><td colspan="2">0 -2 3/2 3/2</td></tr> </table>	0		1/2	1/2	1/3	1/3 0	2/3	1/3 0 1/3	0 -2 3/2 3/2	
0																														
1/3	1/3																													
2/3	1/3 1/3																													
1/3 1/3 1/3																														
0																														
1/2	1/2																													
1/3	1/3 0																													
2/3	1/3 0 1/3																													
0 -1 1 1																														
0																														
1/2	1/2																													
1/3	1/3 0																													
2/3	1/3 0 1/3																													
0 -2 3/2 3/2																														
$T_{3,1}$ order 1	$T_{3,2}$ order 2	$T_{3,3}$ order 3																												

2. Combine (9.18) and (9.19) into the formula ($x = x_0 + kh$)

$$y(x) - y_k = \sum_{j=1}^{\ell} \left(\alpha_{2j}(x) + (-1)^k \beta_{2j}(x) \right) h^{2j} + h^{2\ell+2} E(x, h)$$

for the asymptotic expansion of the Gragg method defined by (9.13a,b).

3. (Stetter 1970). Prove that for every real b (generally between 0 and 1) the method

$$y_1 = y_0 + h \left(b f(x_0, y_0) + (1-b) f(x_1, y_1) \right)$$

$$y_{i+1} = y_{i-1} + h \left((1-b) f(x_{i-1}, y_{i-1}) + 2b f(x_i, y_i) + (1-b) f(x_{i+1}, y_{i+1}) \right)$$

possesses an expansion in powers of h^2 . Prove the same property for the smoothing step

$$S_h(x) = \frac{1}{2} \left(y_{2n} + y_{2n-1} + h(1-b) f(x_{2n-1}, y_{2n-1}) + h b f(x_{2n}, y_{2n}) \right).$$

4. (Stetter 1970). Is the Euler step (9.13a) essential for an h^2 -expansion? Prove that a first order starting procedure

$$y_1 = y_0 + h \Phi(x_0, y_0, h)$$

for (9.13a) produces an h^2 -expansion if the quantities

$y_{-1} = y_0 - h \Phi(x_0, y_0, -h)$, y_0 , and y_1 satisfy (9.13b) for $i = 0$.

5. Study the numerical instability of the extrapolation scheme for the harmonic sequence, i.e., suppose that the entries T_{11} , T_{21} , $T_{31} \dots$ are disturbed with rounding errors ε , $-\varepsilon$, ε, \dots and compute the propagation of these errors into the extrapolation tableau (9.5).

Result. Due to the linearity of the extrapolation scheme, we suppose the T_{ik} equal zero and $\varepsilon = 1$. Then the results for sequence (9.8') are

1.								
-1.	-1.67							
1.	2.60	3.13						
-1.	-3.57	-5.63	-6.21					
1.	4.56	9.13	11.94	12.69				
-1.	-5.55	-13.63	-21.21	-25.35	-26.44			
1.	6.54	19.13	35.01	47.65	54.14	55.82		
-1.	-7.53	-25.63	-54.31	-84.09	-105.64	-116.30	-119.03	
1.	8.53	33.13	80.13	140.14	195.34	232.96	251.10	255.73

hence, for order 18, we lose approximately two digits due to roundoff errors.

6. (Laguerre 1883^{*}). If a_1, a_2, \dots, a_n are distinct positive real numbers and r_1, r_2, \dots, r_n are distinct reals, then

$$A = \begin{pmatrix} a_1^{r_1} & a_1^{r_2} & \dots & a_1^{r_n} \\ a_2^{r_1} & a_2^{r_2} & \dots & a_2^{r_n} \\ \vdots & \vdots & & \vdots \\ a_n^{r_1} & a_n^{r_2} & \dots & a_n^{r_n} \end{pmatrix}$$

is invertible.

Hint (Pólya & Szegő 1925, Vol. II, Abschn. V, Problems 76-77^{*}). Show by induction on n that, if the function $g(t) = \sum_{i=1}^n \alpha_i t^{r_i}$ has n distinct positive zeros, then $g(t) \equiv 0$. By Rolle's theorem the function

$$\frac{d}{dt}(t^{-r_1} g(t)) = \sum_{i=2}^n \alpha_i (r_i - r_1) t^{r_i - r_1 - 1}$$

has $n - 1$ positive distinct zeros and the induction hypothesis can be applied.

^{*} We are grateful to our colleague J. Steinig for these references.

II.10 Numerical Comparisons

The Pleiades seem to be among the first stars mentioned in astronomical literature, appearing in Chinese annals of 2357 B.C. . . .

(R.H. Allen, *Star names, their love and meaning*, 1899, Dover 1963)

If you enjoy fooling around making pictures, instead of typesetting ordinary text, \TeX will be a source of endless frustration/amusement for you, . . .

(D. Knuth, *The \TeX book*, p. 389)

Problems

EULR — Euler’s equation of rotation of a rigid body (“Diese merkwürdig symmetrischen und eleganten Formeln . . .”, A. Sommerfeld 1942, vol. I, § 26.1, Euler 1758)

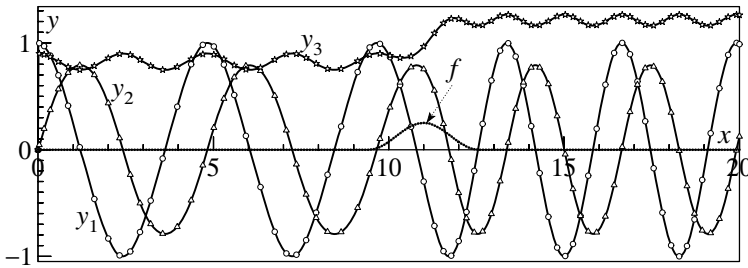


Fig. 10.1. Solutions of Euler’s equations (10.1)

$$\begin{aligned} I_1 y_1' &= (I_2 - I_3) y_2 y_3 \\ I_2 y_2' &= (I_3 - I_1) y_3 y_1 \\ I_3 y_3' &= (I_1 - I_2) y_1 y_2 + f(x) \end{aligned} \quad (10.1)$$

where y_1, y_2, y_3 are the coordinates of $\vec{\omega}$, the rotation vector, and I_1, I_2, I_3 are the principal moments of inertia. The third coordinate has an additional exterior force

$$f(x) = \begin{cases} 0.25 \cdot \sin^2 x & \text{if } 3\pi \leq x \leq 4\pi \\ 0 & \text{otherwise} \end{cases} \quad (10.1')$$

which is discontinuous in its second derivative. We choose the constants and initial values as

$$I_1 = 0.5, \quad I_2 = 2, \quad I_3 = 3, \quad y_1(0) = 1, \quad y_2(0) = 0, \quad y_3(0) = 0.9$$

(see Fig. 10.1) and check the numerical precision at the output points

$$x_{\text{end}} = 10 \quad \text{and} \quad x_{\text{end}} = 20 .$$

AREN — the Arenstorf orbit (0.1) for the restricted three body problem with initial values (0.2) integrated over one period $0 \leq x \leq x_{\text{end}}$ (see Fig. 0.1). The precision is checked at the endpoint, here the solution is most sensitive to errors of the initial phase.

LRNZ — the solution of the Saltzman-Lorenz equations (I.16.17) displayed in Fig. I.16.8, i.e., with constants and initial values

$$\sigma = 10, \quad r = 28, \quad b = \frac{8}{3}, \quad y_1(0) = -8, \quad y_2(0) = 8, \quad y_3(0) = 27 . \quad (10.2)$$

The solution is, for large values of x , *extremely* sensitive to the errors of the first integration steps (see Fig. I.16.10 and its discussion). For example, at $x = 50$ the numerical solution becomes totally wrong, even if the computations are performed in quadruple precision with $\text{Tot} = 10^{-20}$. Hence the numerical results of *all* methods would be equally useless and no comparison makes any sense. Therefore we choose

$$x_{\text{end}} = 16$$

and check the numerical solution at this point. Even here, all computations with $\text{Tot} \geq 10^{-7}$, say, fall into a chaotic cloud of meaningless results (see Fig. 10.5).

PLEI — a celestial mechanics problem (which we call “the Pleiades”): seven stars in the plane with coordinates x_i, y_i and masses $m_i = i$ ($i = 1, \dots, 7$):

$$\begin{aligned} x_i'' &= \sum_{j \neq i} m_j (x_j - x_i) / r_{ij} \\ y_i'' &= \sum_{j \neq i} m_j (y_j - y_i) / r_{ij} \end{aligned} \quad (10.3)$$

where

$$r_{ij} = ((x_i - x_j)^2 + (y_i - y_j)^2)^{3/2}, \quad i, j = 1, \dots, 7.$$

The initial values are

$$\begin{aligned} x_1(0) &= 3, & x_2(0) &= 3, & x_3(0) &= -1, & x_4(0) &= -3, \\ x_5(0) &= 2, & x_6(0) &= -2, & x_7(0) &= 2, \\ y_1(0) &= 3, & y_2(0) &= -3, & y_3(0) &= 2, & y_4(0) &= 0, \\ y_5(0) &= 0, & y_6(0) &= -4, & y_7(0) &= 4, \\ x_i'(0) &= y_i'(0) = 0, & \text{for all } i & \text{with the exception of} \\ x_6'(0) &= 1.75, & x_7'(0) &= -1.5, & y_4'(0) &= -1.25, & y_5'(0) &= 1, \end{aligned} \quad (10.4)$$

and we integrate for $0 \leq t \leq t_{\text{end}} = 3$. Fig. 10.2a represents the movement of these 7 bodies in phase coordinates. The initial value is marked by an “i”, the final value at $t = t_{\text{end}}$ is marked by an “f”. Between these points, 19 time-equidistant output points are plotted and connected by a dense output formula. There occur several quasi-collisions which are displayed in Table 10.1.

Table 10.1. Quasi-collisions in the PLEI problem

Body 1	1	1	3	1	2	5
Body 2	7	3	5	7	6	7
r_{ij}^2	0.0129	0.0193	0.0031	0.0011	0.1005	0.0700
time	1.23	1.46	1.63	1.68	1.94	2.14

The resulting violent shapes of the derivatives $x'_i(t), y'_i(t)$ are displayed in Fig. 10.2b and show that automatic step size control is essential for this example.

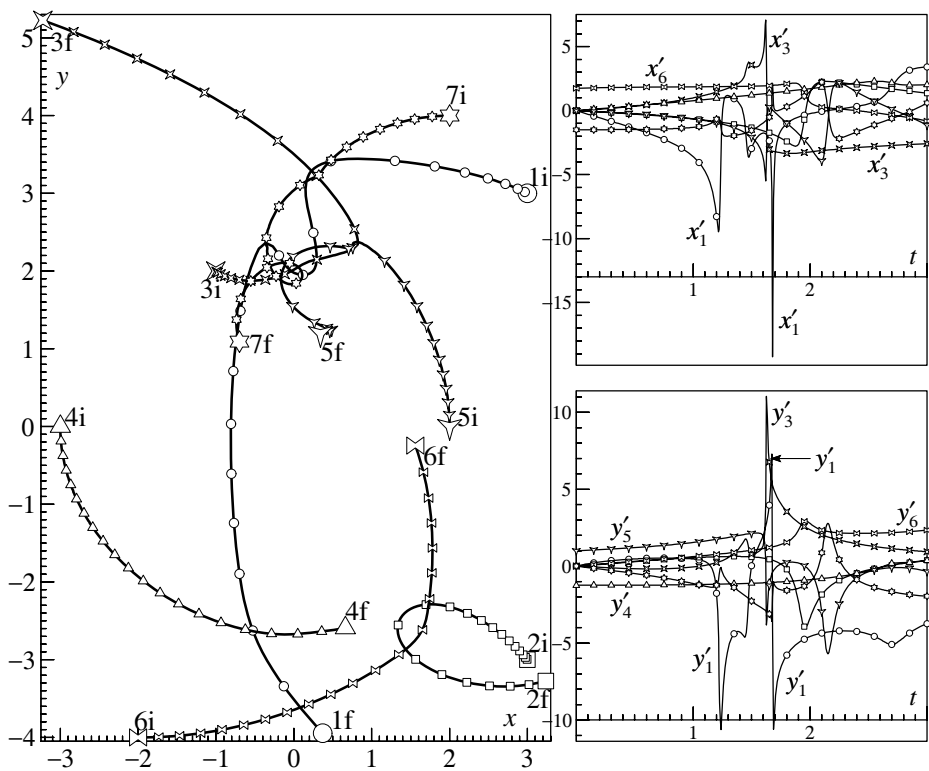


Fig. 10.2a. Solutions of (10.3)

Fig. 10.2b. Speeds

ROPE — the movement of a hanging rope (see Fig. 10.3a) of length 1 under gravitation and under the influence of a horizontal force

$$F_y(t) = \left(\frac{1}{\cosh(4t - 2.5)} \right)^4 \quad (10.5a)$$

acting at the point $s = 0.75$ as well as a vertical force

$$F_x(t) = 0.4 \quad (10.5b)$$

acting at the endpoint $s = 1$.

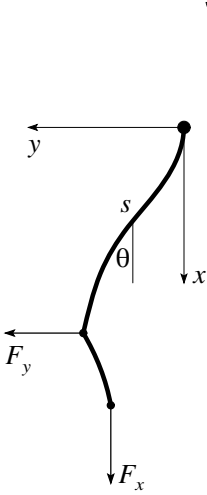


Fig. 10.3a. Hanging rope

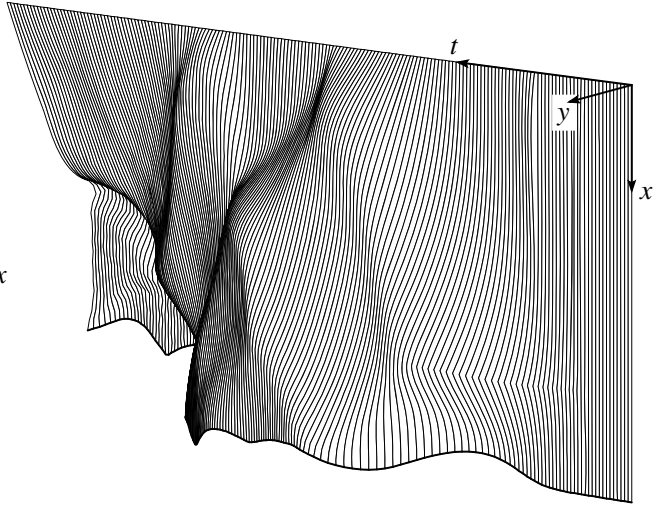


Fig. 10.3b. Solution for $0 \leq t \leq 3.723$.

If this problem is discretized, then Lagrange theory (see (I.6.18); see also Exercises IV.1.2 and IV.1.4 of Volume II) leads to the following equations for the unknown angles θ_k :

$$\sum_{k=1}^n a_{lk} \ddot{\theta}_k = - \sum_{k=1}^n b_{lk} \dot{\theta}_k^2 - n \left(n + \frac{1}{2} - l \right) \sin \theta_l - n^2 \sin \theta_l \cdot F_x(t) + \begin{cases} n^2 \cos \theta_l \cdot F_y(t) & \text{if } l \leq 3n/4 \\ 0 & \text{if } l > 3n/4, \end{cases} \quad l = 1, \dots, n \quad (10.6)$$

where

$$a_{lk} = g_{lk} \cos(\theta_l - \theta_k), \quad b_{lk} = g_{lk} \sin(\theta_l - \theta_k), \quad g_{lk} = n + \frac{1}{2} - \max(l, k). \quad (10.7)$$

We choose

$$n = 40, \quad \theta_l(0) = \dot{\theta}_l(0) = 0, \quad 0 \leq t \leq 3.723. \quad (10.8)$$

The resulting system is of dimension 80. The special structure of G^{-1} (see (IV.1.16–18) of Volume II) allows one to evaluate $\dot{\theta}_l$ with the following algorithm:

- a) Let $v_l = -n(n + \frac{1}{2} - l) \sin \theta_l - n^2 \sin \theta_l \cdot F_x + \begin{cases} n^2 \cos \theta_l \cdot F_y \\ 0 \end{cases}$
- b) Compute $w = Dv + \dot{\theta}^2$,
- c) Solve the tridiagonal system $Cu = w$,
- d) Compute $\ddot{\theta} = Cv + Du$,

where

$$C = \begin{pmatrix} 1 & -\cos(\theta_1 - \theta_2) & & & \\ -\cos(\theta_2 - \theta_1) & 2 & -\cos(\theta_2 - \theta_3) & & \\ & -\cos(\theta_3 - \theta_2) & \ddots & \ddots & \\ & & \ddots & 2 & -\cos(\theta_{n-1} - \theta_n) \\ & & & -\cos(\theta_n - \theta_{n-1}) & 3 \end{pmatrix} \quad (10.9)$$

$$D = \begin{pmatrix} 0 & -\sin(\theta_1 - \theta_2) & & & \\ -\sin(\theta_2 - \theta_1) & 0 & -\sin(\theta_2 - \theta_3) & & \\ & -\sin(\theta_3 - \theta_2) & \ddots & \ddots & \\ & & \ddots & 0 & -\sin(\theta_{n-1} - \theta_n) \\ & & & -\sin(\theta_n - \theta_{n-1}) & 0 \end{pmatrix}.$$

BRUS — the reaction-diffusion equation (Brusselator with diffusion)

$$\begin{aligned} \frac{\partial u}{\partial t} &= 1 + u^2 v - 4.4u + \alpha \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) \\ \frac{\partial v}{\partial t} &= 3.4u - u^2 v + \alpha \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right) \end{aligned} \quad (10.10)$$

for $0 \leq x \leq 1$, $0 \leq y \leq 1$, $t \geq 0$, $\alpha = 2 \cdot 10^{-3}$ together with the Neumann boundary conditions

$$\frac{\partial u}{\partial \mathbf{n}} = 0, \quad \frac{\partial v}{\partial \mathbf{n}} = 0, \quad (10.11)$$

and the initial conditions

$$u(x, y, 0) = 0.5 + y, \quad v(x, y, 0) = 1 + 5x. \quad (10.12)$$

By the method of lines (cf. Section I.6) this problem becomes a system of ordinary differential equations. We put

$$x_i = \frac{i-1}{N-1}, \quad y_j = \frac{j-1}{N-1}, \quad i, j = 1, \dots, N$$

and define

$$U_{ij}(t) = u(x_i, y_j, t), \quad V_{ij}(t) = v(x_i, y_j, t). \quad (10.13)$$

Discretizing the derivatives in (10.10) with respect to the space variables we obtain for $i, j = 1, \dots, N$

$$\begin{aligned} U'_{ij} &= 1 + U_{ij}^2 V_{ij} - 4.4 U_{ij} + \alpha(N-1)^2 (U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} - 4U_{ij}) \\ V'_{ij} &= 3.4 U_{ij} - U_{ij}^2 V_{ij} + \alpha(N-1)^2 (V_{i+1,j} + V_{i-1,j} + V_{i,j+1} + V_{i,j-1} - 4V_{ij}), \end{aligned} \quad (10.14)$$

an ODE of dimension $2N^2$. Because of the boundary condition (10.11) we have

$$U_{0,j} = U_{2,j}, \quad U_{N+1,j} = U_{N-1,j}, \quad U_{i,0} = U_{i,2}, \quad U_{i,N+1} = U_{i,N-1}$$

and similarly for the V_{ij} -quantities. We choose $N = 21$ so that the system is of dimension 882 and check the numerical solutions at the output point $t_{\text{end}} = 7.5$. The solution of (10.14) (in the (x, y) -space) is represented in Fig. 10.4a and Fig. 10.4b for u and v respectively.

Performance of the Codes

Several codes were applied to each of the test problems with $Tol = 10^{-3}$, $Tol = 10^{-3-1/8}$, $Tol = 10^{-3-2/8}$, $Tol = 10^{-3-3/8}$, \dots (for the large problems with $Tol = 10^{-3}$, $Tol = 10^{-3-1/4}$, $Tol = 10^{-3-2/4}$, \dots) up to, in general, $Tol = 10^{-14}$, then the numerical result at the output points were compared with an “exact solution” (computed very precisely in quadruple precision). Each of these results then corresponds to one point of Fig. 10.5, where this precision is compared (in double logarithmic scale) to the number of function evaluations. The “integer” tolerances 10^{-3} , 10^{-4} , 10^{-5} , \dots are distinguishable as enlarged symbols. All codes were applied with complete “standard” parameter settings and were not at all “tuned” to these particular problems.

A comparison of the *computing time* (instead of the number of function evaluations) gave no significant difference. Therefore, only one representative of the small problems (LRNZ) and one large problem (BRUS) are displayed in Fig. 10.6. All computations have been performed in REAL*8 ($U_{\text{round}} = 1.11 \cdot 10^{-16}$) on a Sun Workstation (SunBlade 100).

The codes used are the following:

RKF45 — symbol \star — a product of Shampine and Watts’ programming art based on Fehlberg’s pair of orders 4 and 5 (Table 5.1). The method is used in the “local extrapolation mode”, i.e., the numerical solution is advanced with the 5th order result. The code is usually, except for low precision, the slowest of all, which is explained by its low order. The results of the “time”-picture Fig. 10.6 for this

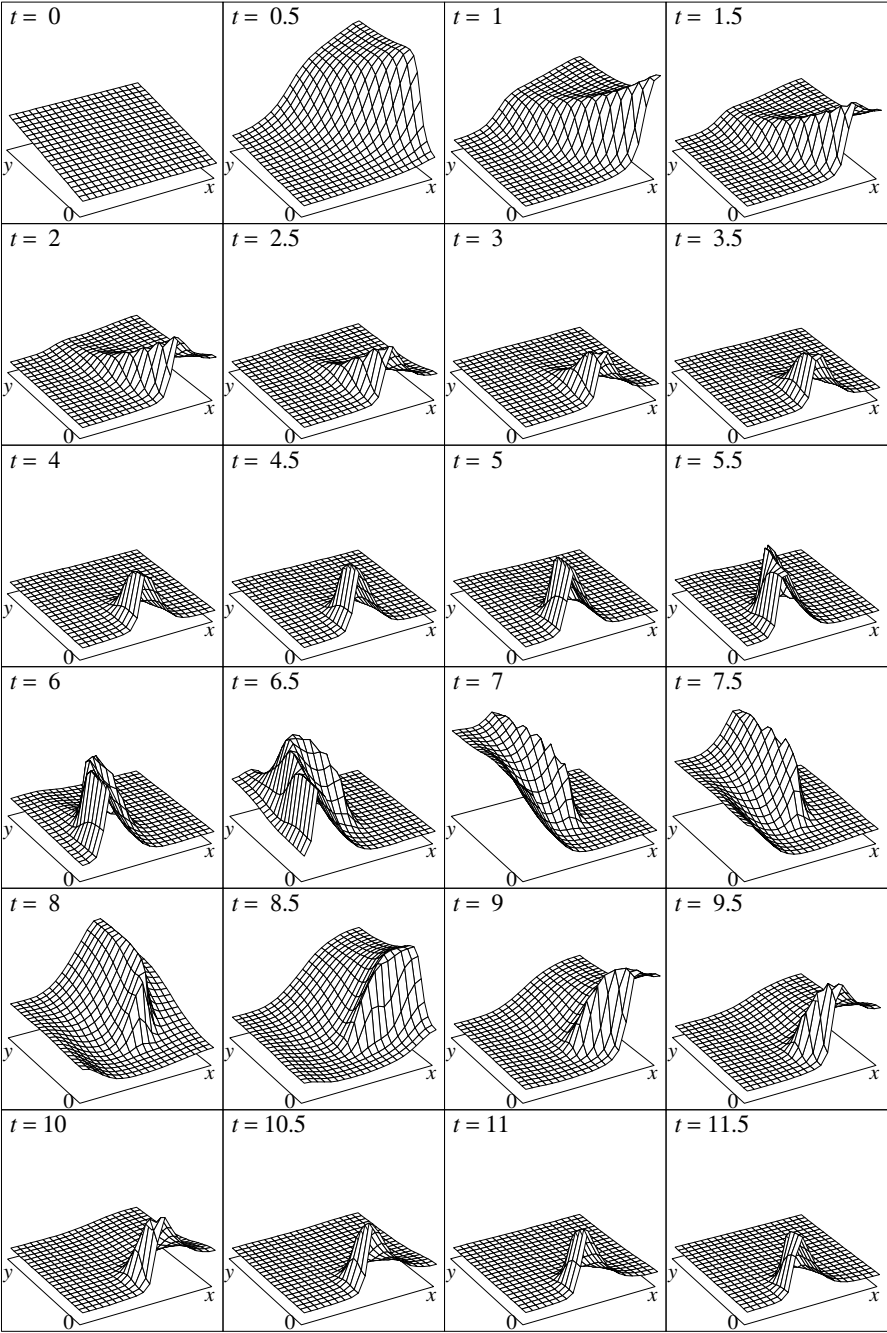


Fig. 10.4a. Solution $u(x, y, t)$ for the BRUS problem

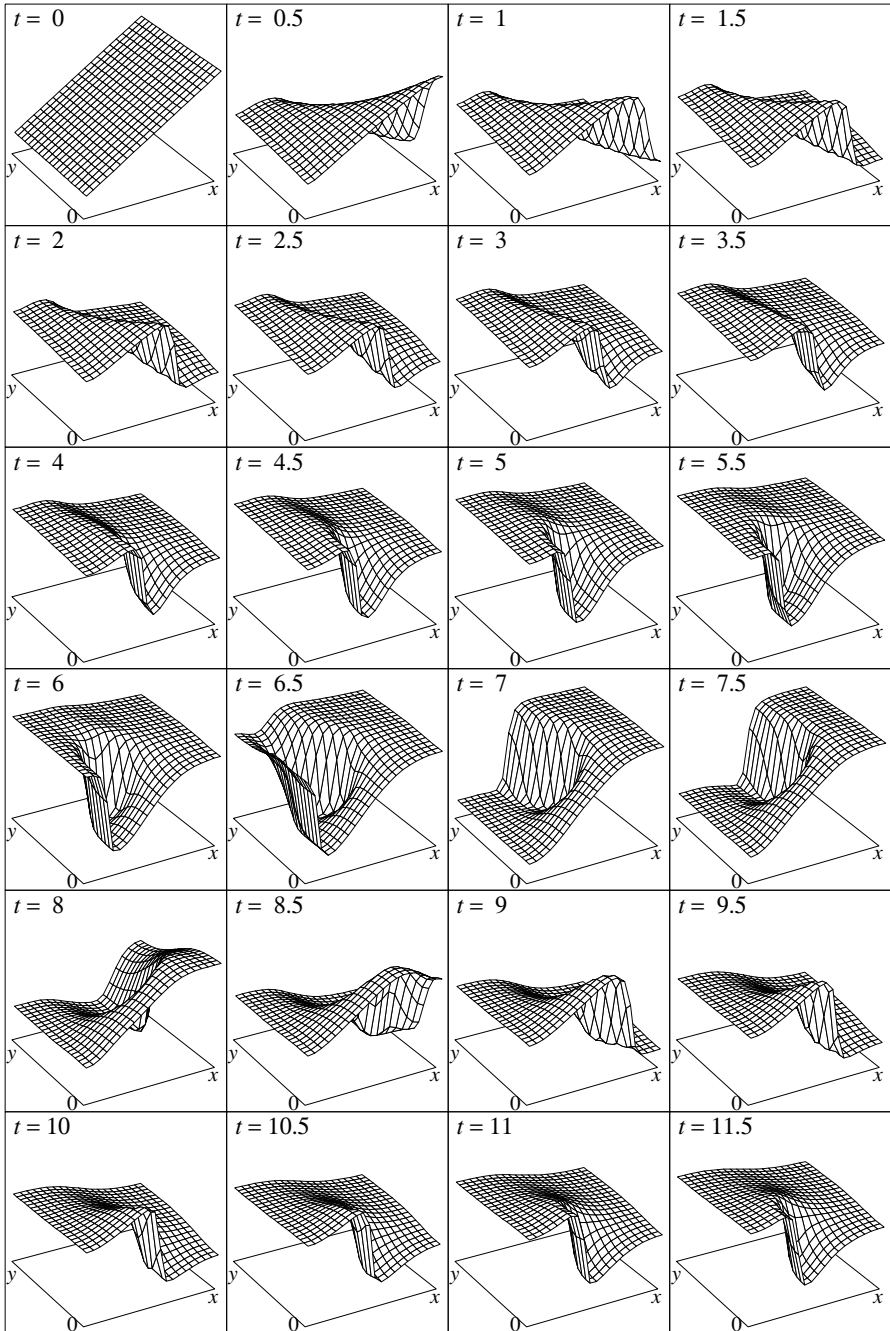


Fig. 10.4b. Solution $v(x, y, t)$ for the BRUS problem

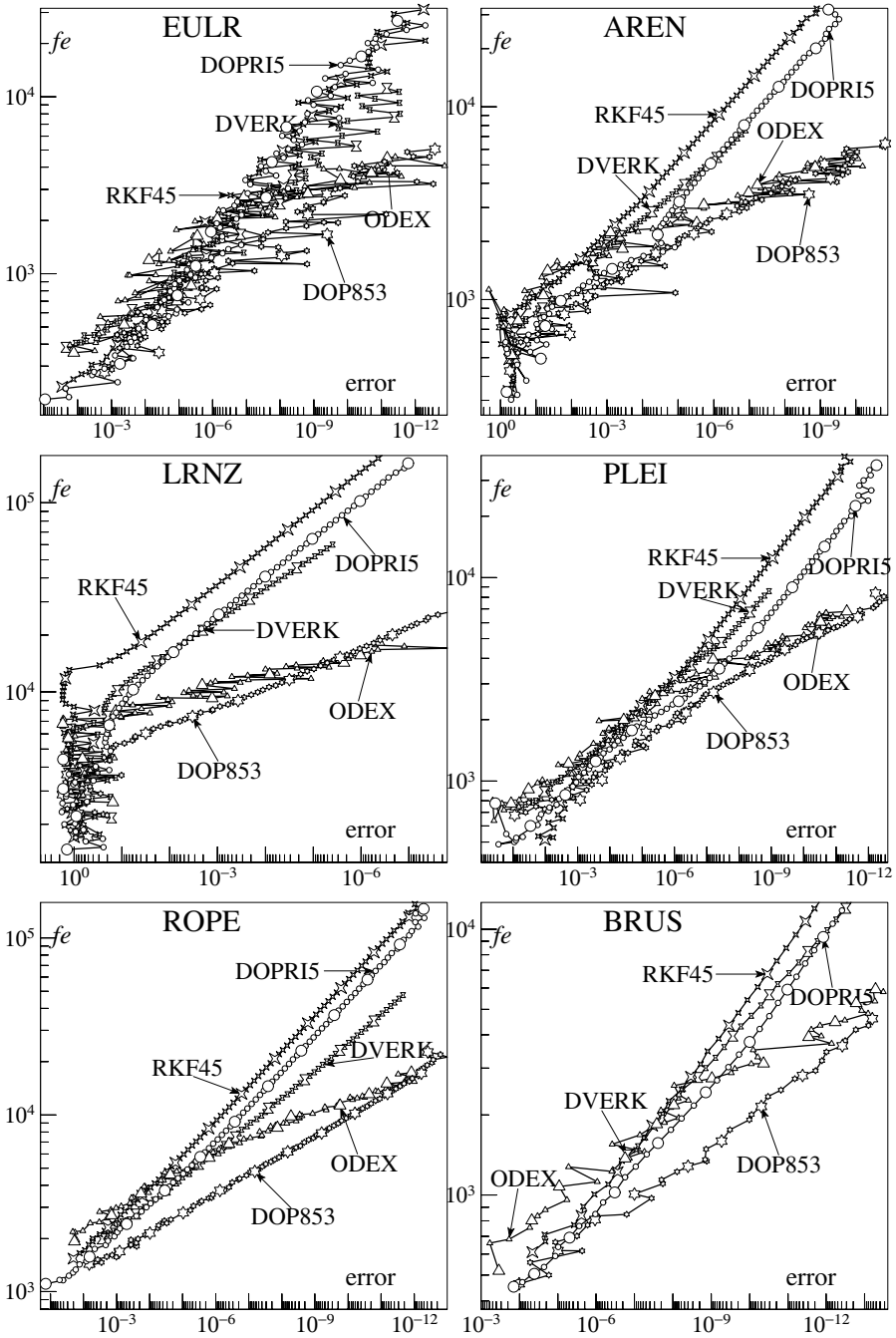


Fig. 10.5. Precision versus function calls

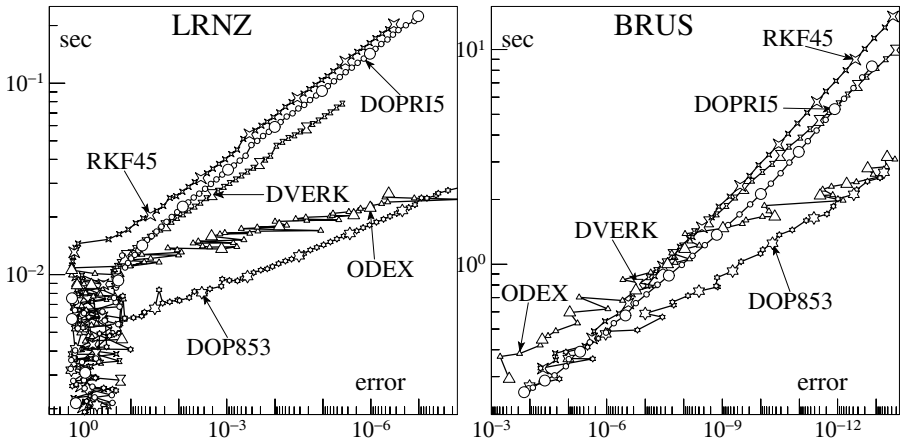


Fig. 10.6. Precision versus computing time

code are relatively better than those on the “function calls” front (Fig. 10.5). This indicates that the code has particularly small overhead.

DOPRI5 — symbol \bigcirc — the method of Dormand & Prince of order 5 with embedded error estimator of order 4 (see Table 5.2). The code is explained in the Appendix. The method has precisely the same order as that used in RKF45, but the error constants are much more optimized. Therefore the “error curves” in Fig. 10.5 are nicely parallel to those of RKF45, but appear translated to the side of higher precision. One usually gains between a half and one digit of numerical precision for comparable numerical work. The code performs specially well between $Tol = 10^{-3}$ and $Tol = 10^{-8}$ in the AREN problem. This is simply due to an accidental sign change of the error for the most sensitive solution component.

DVERK — symbol Σ — this widely known code implements Verner’s 6th order method of Table 5.4 and was written by Hull, Enright & Jackson. It has been included in the IMSL library for many years and the source code is available through na-net. The corresponding error curves in Fig. 10.5 appear to be less steep than those of DOPRI5, which illustrates the higher order of the method. However, the error constants seem to be less optimal so that this code surpasses the performance of DOPRI5 only for very stringent tolerances. It is significantly better than DOPRI5 solely in problems EULR and ROPE. The code, as it was, failed at the BRUS problem for $Tol = 10^{-3}$ and $Tol = 10^{-4}$. Therefore these computations were started with $Tol = 10^{-5}$.

DOP853 — symbol \star — is the method of Dormand & Prince of order 8 explained in Section II.5 (formulas (5.20) – (5.30), see Appendix). The 6th order error estimator (5.29), (5.30) has been replaced by a 5th order estimator with 3rd order correction (see below). This was necessary to make the code robust for the

EULR problem. The code works perfectly for all problems and nearly all tolerances. Whenever more than 3 or 4 digits are desired, this method seems to be highly recommendable. The most astonishing fact is that its use was never disastrous, even not for $Tol = 10^{-3}$.

ODEX — symbol \triangle — is an extrapolation code based on the Gragg-Bulirsch-Stoer algorithm with harmonic step number sequence (see Appendix). This method, which allows arbitrary high orders (in the standard version of the code limited to $p \leq 18$) is of course predestined for computations with high precision. The more stringent Tol is, the higher the used order becomes, the less steep the error curve is. This can best be observed in the picture for the ROPE problem. Finally, for $Tol \approx 10^{-12}$, the code surpasses the values of DOP853. As can be seen in Fig. 10.6, the code loses slightly on the “time”-front. This is due to the increased overhead of the extrapolation scheme.

The numerical results of ODEX behave very similarly to those of DIFEX1 (Deuffhard 1983).

A “Stretched” Error Estimator for DOP853

In preliminary stages of our numerical tests we had written a code “DOPR86” based on the method of order 8 of Dormand & Prince with the 6th order error estimator described in Section II.5. For most problems the results were excellent. However, there are some situations in which the error control of DOPR86 did not work safely:

When applied to the BRUS problem with $Tol = 10^{-3}$ or $Tol = 10^{-4}$ the code stopped with an overflow message. The reason was the following: when the step size is too large, the internal stages are too far away from the solution and their modulus increases at each stage (e.g., by a factor 10^5 between stage 11 and stage 12). Due to the fact that $\hat{b}_{12} = b_{12}$ (see (5.30) (5.26) and (5.25b)) the difference $\hat{y}_1 - y_1$ is not influenced by the last stage and is smaller (by a factor of 10^5) than the modulus of y_1 . Hence, the error estimator scaled by (4.10) is $\leq 10^{-5}$ and a completely wrong step will be accepted.

The code DOPR86 also had severe difficulties when applied to problems with discontinuities such as EULR. The worst results were obtained for the problem

$$\begin{aligned} y_1' &= y_2 y_3 & y_1(0) &= 0 \\ y_2' &= -y_3 y_1 & y_2(0) &= 1 \\ y_3' &= -0.51 \cdot y_1 y_2 + f(x) & y_3(0) &= 1 \end{aligned} \quad (10.15)$$

where $f(x)$, given in (10.1'), has a discontinuous second derivative. The results for this problem and the code DOPR86 for very many different Tol values ($Tol = 10^{-3}, 10^{-3-1/24}, 10^{-3-2/24}, \dots, 10^{-14}$) are displayed in Fig. 10.7. There,

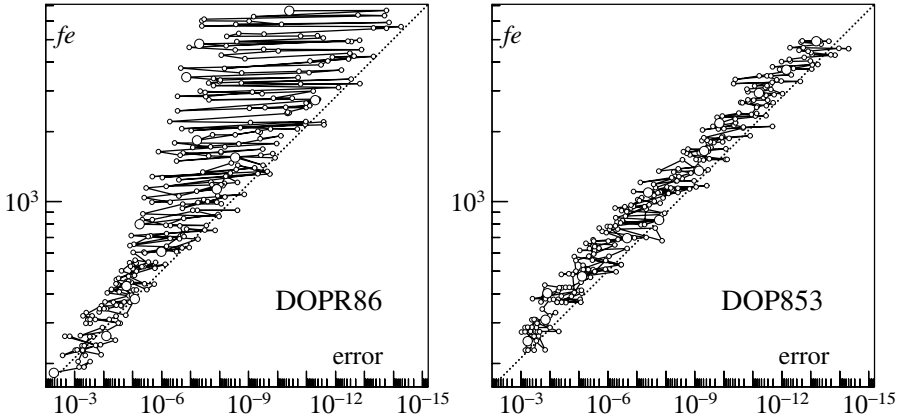


Fig. 10.7. Performances of DOPR86 and DOP853 at (10.15)

the (dotted) diagonal is of exact slope $1/8$ and represents the theoretical convergence speed of the method of order 8. It can be observed that this convergence is well attained by *some* results, but others lose precision of up to 8 digits from the desired tolerance. We explain this disappointing behaviour by the fact that $\hat{b}_{12} = b_{12}$ and that the 12th stage is the only one where the function is evaluated at the end-point of the step. Whenever the discontinuity of f'' is by accident slightly to the left of a grid point, the error estimator ignores it and the code reports a wrong value.

Unfortunately, the basic 8th order method does not possess a 6th order embedding with $\hat{b}_{12} \neq b_{12}$ (unless additional function evaluations are used). Therefore, we decided to construct a 5th order approximation \hat{y}_1 . It can be obtained by taking $\hat{b}_6, \hat{b}_7, \hat{b}_{12}$ as free parameters, e.g.,

$$\hat{b}_6 = b_6/2 + 1, \quad \hat{b}_7 = b_7/2 + 0.45, \quad \hat{b}_{12} = b_{12}/2,$$

by putting $\hat{b}_2 = \hat{b}_3 = \hat{b}_4 = \hat{b}_5 = 0$ and by determining the remaining coefficients such that this quadrature formula has order 5. Due to the simplifying assumptions (5.20) all conditions for order 5 are then satisfied. In order to prevent a serious *over-estimation* of the error, we consider a second embedded method \tilde{y}_1 of order 3 based on the nodes $c_1 = 0$, c_9 and $c_{12} = 1$ so that two error estimators

$$err_5 = \|\hat{y}_1 - y_1\| = \mathcal{O}(h^6), \quad err_3 = \|\tilde{y}_1 - y_1\| = \mathcal{O}(h^4) \quad (10.16)$$

are available. Similarly to a procedure which is common for quadrature formulas (R. Piessens, E. de Doncker-Kapenga, C.W. Überhuber & D.K. Kahaner 1983, Berntsen & Espelid 1991) we consider

$$err = err_5 \cdot \frac{err_5}{\sqrt{err_5^2 + 0.01 \cdot err_3^2}} = \mathcal{O}(h^8) \quad (10.17)$$

as error estimator. It behaves asymptotically like the global error of the method. The corresponding code DOP853 gives satisfactory results for all the above problems (see right picture in Fig. 10.7).

Effect of Step-Number Sequence in ODEX

We also study the influence of the different step-number sequences to the performance of the extrapolation code ODEX. Fig. 10.8 presents two examples of this study, a small problem (AREN) and a large problem (ROPE). The used sequences are

HARMONIC — symbol \bigcirc — the harmonic sequence (9.8') which is the standard choice in ODEX;

MOD4 — symbol \triangle — the sequence $\{2, 6, 10, 14, 18, \dots\}$ (see (9.35)) which allowed the construction of high-order dense output;

BULIRSCH — symbol \square — the Bulirsch sequence (9.7');

ROMBERG — symbol \diamond — the Romberg sequence (9.6');

DNSECTRL — symbol \star — the error control for the MOD4 sequence taking into account the interpolation error of the dense output solution (9.42). This is included only in the small problem, since (complete) dense output on large problems would need too much memory.

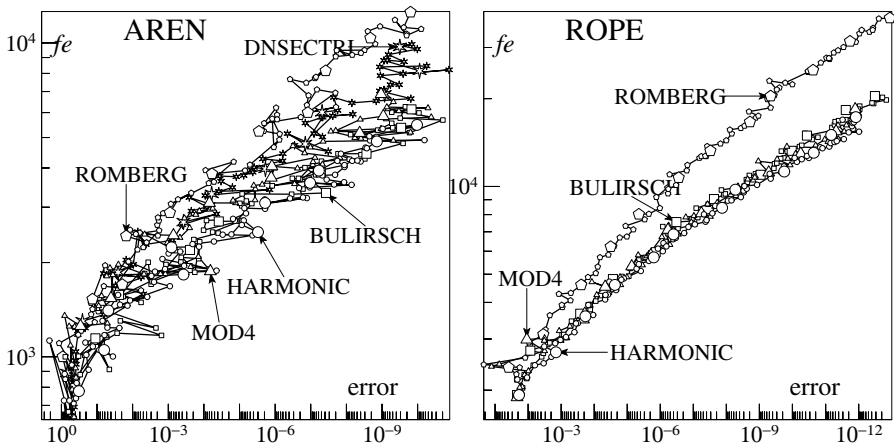


Fig. 10.8. Effect of step-number sequences in ODEX

Discussion. With the exception of the clear inferiority of the Romberg sequence, especially for high precision, and a certain price to be paid for the dense output error control, there is not much difference between the first three sequences. Although the harmonic sequence appears to be slightly superior, the difference is statistically not very significant.

II.11 Parallel Methods

We suppose that we have a computer with a number of arithmetic processors capable of simultaneous operation and seek to devise parallel integration algorithms for execution on such a computer.

(W.L. Miranker & W. Liniger 1967)

“PARALYSING ODES” (K. Burrage, talk in Helsinki 1990)

Parallel machines are computers with more than one processor and this facility might help us to speed up the computations in ordinary differential equations. This is particularly interesting for very large problems, for very costly function evaluation, or for fast real-time simulations. A second motivation is the desire to make a code, with the help of parallel computations, not necessarily faster but more robust and reliable.

Early attempts for finding parallel methods are Nievergelt (1964) and Miranker & Liniger (1967). See also the survey papers Miranker (1971) and Jackson (1991).

We distinguish today essentially between two types of parallel architectures:

SIMD (single instruction multiple data): all processors execute the same instructions with possibly different input data.

MIMD (multiple instruction multiple data): the different processors can act independently.

The exploitation of parallelism for an ordinary differential equation

$$y' = f(x, y), \quad y(x_0) = y_0 \quad (11.1)$$

can be classified into two main categories (Gear 1987, 1988):

Parallelism across the system. Often the problem itself offers more or less trivial applications for parallelism, e.g.,

- > if several solutions are required for various initial or parameter values;
- > if the right-hand side of (11.1) is very costly, but structured in such a way that the computation of *one* function evaluation can be split efficiently across the various processors;
- > space discretizations of partial differential equations (such as the Brusselator problem (10.14)) whose function evaluation can be done simultaneously for all components on an SIMD machine with thousands of processors;
- > the solution of boundary value problems with the multiple shooting method (see Section I.15) where all computations on the various sub-intervals can be done in parallel;

- > doing all the high-dimensional linear algebra in the Runge-Kutta method (11.2) in parallel;
- > parallelism in the linear algebra for Newton's method for *implicit* Runge-Kutta methods (see Section IV.8).

These types of parallelism, of course, depend strongly on the problem and on the type of the computer.

Parallelism across the method. This is problem-independent and means that, due to a special structure of the method, several function values can be evaluated in parallel within one integration step. This will be discussed in this section in more detail.

Parallel Runge-Kutta Methods

... it seems that *explicit* Runge-Kutta methods are not facilitated much by parallelism at the method level.

(Iserles & Nørsett 1990)

Consider an explicit Runge-Kutta method

$$k_i = f\left(x_0 + c_i h, y_0 + h \sum_{j=1}^{i-1} a_{ij} k_j\right), \quad i = 1, \dots, s$$

$$y_1 = y_0 + h \sum_{i=1}^s b_i k_i.$$
(11.2)

Suppose, for example, that the coefficients have the zero-pattern indicated in Fig. 11.1.

0				
×	×			
×	×	0		
×	×	×	×	
	×	×	×	×

Fig. 11.1. Parallel method

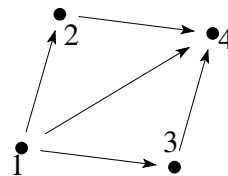


Fig. 11.2. Production graph

Each arrow in the corresponding “production graph” G (Fig. 11.2), pointing from vertex “ i ” to vertex “ j ”, stands for a non-zero a_{ji} . Here the vertices 2 and 3 are independent and can be evaluated in parallel. We call the number of vertices in the longest chain of successive arrows (here 3) the *number of sequential function evaluations* σ .

In general, if the Runge-Kutta matrix A can be partitioned (possibly after a permutation of the stages) as

$$A = \begin{pmatrix} 0 & & & & \\ A_{21} & 0 & & & \\ A_{31} & A_{32} & 0 & & \\ \vdots & \vdots & & \ddots & \\ A_{\sigma 1} & A_{\sigma 2} & \dots & A_{\sigma, \sigma-1} & 0 \end{pmatrix}, \quad (11.3)$$

where A_{ij} is a matrix of size $\mu_i \times \mu_j$, then the derivatives k_1, \dots, k_{μ_1} as well as $k_{\mu_1+1}, \dots, k_{\mu_1+\mu_2}$, and so on, can be computed in parallel and one step of the method is executed in σ sequential function evaluations (if $\mu = \max_i \mu_i$ processors are at disposal). The following theorem is a severe restriction on parallel methods. It appeared in hand-written notes by K. Jackson & S. Nørsett around 1986. For a publication see Jackson & Nørsett (1992) and Iserles & Nørsett (1990).

Theorem 11.1. *For an explicit Runge-Kutta method with σ sequential stages the order p satisfies*

$$p \leq \sigma, \quad (11.4)$$

for any number μ of available processors.

Proof. Each non-zero term of the expressions $\Phi_i(t)$ for the “tall” trees $t_{21}, t_{32}, t_{44}, t_{59}, \dots$ (see Table 2.2 and Definition 2.9) $\sum a_{ij} a_{jk} a_{k\ell} a_{\ell m} \dots$ corresponds to a connected chain of arrows in the production graph. Since their length is limited by σ , these terms are all zero for $\varrho(t) > \sigma$. \square

Methods with $p = \sigma$ will be called *P-optimal methods*. The Runge-Kutta methods of Section II.1 for $p \leq 4$ are all P-optimal. Only for $p > 4$ does the subsequent construction of P-optimal methods allow one to increase the order with the help of parallelism.

Remark. The fact that the “stability function” (see Section IV.2) of an explicit parallel Runge-Kutta method is a polynomial of degree $\leq \sigma$ allows a second proof of Theorem 11.1. Further, P-optimal methods all have the same stability function $1 + z + z^2/2! + \dots + z^\sigma/\sigma!$.

Parallel Iterated Runge-Kutta Methods

One possibility of constructing P-optimal methods is by fixed point iteration. Consider an arbitrary (explicit or implicit) Runge-Kutta method with coefficients

$$c = (c_1, \dots, c_s)^T, \quad A = (a_{ij})_{i,j=1}^s, \quad b^T = (b_1, \dots, b_s)$$

and define \hat{y}_1 by

$$\begin{aligned} k_i^{(0)} &= 0 \\ k_i^{(\ell)} &= f\left(x_0 + c_i h, y_0 + h \sum_{j=1}^s a_{ij} k_j^{(\ell-1)}\right), \quad \ell = 1, \dots, \sigma \\ \hat{y}_1 &= y_0 + h \sum_{i=1}^s b_i k_i^{(\sigma)}. \end{aligned} \quad (11.5)$$

This algorithm can be interpreted as an explicit Runge-Kutta method with scheme

$$\begin{array}{c|cccccc} 0 & 0 & & & & \\ c & A & 0 & & & \\ c & 0 & A & 0 & & \\ \vdots & \vdots & & \ddots & \ddots & \\ c & 0 & \dots & 0 & A & 0 \\ \hline & 0 & \dots & 0 & 0 & b^T \end{array} \quad (11.6)$$

It has σ sequential stages if s processors are available. To compute its order we use a Lipschitz condition for $f(x, y)$ and obtain

$$\max_i \|k_i^{(\ell)} - k_i\| \leq Ch \cdot \max_i \|k_i^{(\ell-1)} - k_i\|$$

where k_i are the stage-vectors of the basic method. Since $k_i^{(0)} - k_i = \mathcal{O}(1)$ this implies $k_i^{(\sigma)} - k_i = \mathcal{O}(h^\sigma)$ and consequently the difference to the solution of the basic method satisfies $\hat{y}_1 - y_1 = \mathcal{O}(h^{\sigma+1})$.

Theorem 11.2. *The parallel iterated Runge-Kutta method (11.5) is of order*

$$p = \min(p_0, \sigma), \quad (11.7)$$

if p_0 denotes the order of the basic method.

Proof. The statement follows from

$$\hat{y}_1 - y(x_0 + h) = \hat{y}_1 - y_1 + y_1 - y(x_0 + h) = \mathcal{O}(h^{\sigma+1}) + \mathcal{O}(h^{p_0+1}). \quad \square$$

This theorem shows that the choice $\sigma = p_0$ in (11.5) yields P-optimal explicit Runge-Kutta methods (i.e., $\sigma = p$). If we take as basic method the s -stage collocation method based on the Gaussian quadrature ($p_0 = 2s$) then we obtain a method of order $p = 2s$ which is P-optimal on s processors. P.J. van der Houwen

& B.P. Sommeijer (1990) have done extensive numerical experiments with this method.

Extrapolation Methods

It turns out that the GBS-algorithm (Section II.9) without smoothing step is also P-optimal. Indeed, all the values T_{j1} can be computed independently of each other. If we choose the step number sequence $\{2, 4, 6, 8, 10, 12, \dots\}$ then the computation of T_{k1} requires $2k$ sequential function evaluations. Hence, if k processors are available (one for each T_{j1}), the numerical approximation T_{kk} , which is of order $p=2k$, can be computed with $\sigma=2k$ sequential stages. When the processors are of type MIMD we can compute T_{11} and $T_{k-1,1}$ on one processor ($2+2(k-1)=2k$ function evaluations). Similarly, T_{21} and $T_{k-2,1}$ occupy another processor, etc. In this way, the number of necessary processors is reduced by a factor close to 2 without increasing the number of sequential stages.

The order and step size strategy, discussed in Section II.9, should, of course, be adapted for an implementation on parallel computers. The “hope for convergence in line $k+1$ ” no longer makes sense because this part of the algorithm is now as costly as the whole step. Similarly, there is no reason to accept already $T_{k-1,k-1}$ as numerical approximation, because T_{kk} is computed on the same time level as $T_{k-1,k-1}$. Moreover, the numbers A_k of (9.25) should be replaced by $A_k = n_k$ which will in general increase the order used by the code.

Increasing Reliability

... using parallelism to improve *reliability* and *functionality*
rather than efficiency. (W.H. Enright & D.J. Higham 1991)

For a given Runge-Kutta method parallel computation can be used to give a reliable error estimate or an accurate dense output. This has been advocated by Enright & Higham (1991) and will be the subject of this subsection.

Consider a Runge-Kutta method of order p , choose distinct numbers $0 = \sigma_0 < \sigma_1 < \dots < \sigma_k = 1$ and apply the Runge-Kutta method in parallel with step sizes $\sigma_1 h, \dots, \sigma_{k-1} h, \sigma_k h = h$. This gives approximations

$$y_{\sigma_i} \approx y(x_0 + \sigma_i h) . \quad (11.8)$$

Then compute $f(x_0 + \sigma_i h, y_{\sigma_i})$ and do Hermite interpolation with the values

$$y_{\sigma_i}, h f(x_0 + \sigma_i h, y_{\sigma_i}), \quad i = 0, 1, \dots, k, \quad (11.9)$$

i.e., compute

$$u(\theta) = \sum_{i=0}^k v_i(\theta) y_{\sigma_i} + h \sum_{i=0}^k w_i(\theta) f(x_0 + \sigma_i h, y_{\sigma_i}) \quad (11.10)$$

where $v_i(\theta)$ and $w_i(\theta)$ are the scalar polynomials

$$\left. \begin{aligned} v_i(\theta) &= \ell_i^2(\theta) \cdot (1 - 2\ell_i'(\sigma_i)(\theta - \sigma_i)) \\ w_i(\theta) &= \ell_i^2(\theta) \cdot (\theta - \sigma_i) \end{aligned} \right\} \text{ with } \ell_i(\theta) = \prod_{\substack{j=0 \\ j \neq i}}^k \frac{(\theta - \sigma_j)}{(\sigma_i - \sigma_j)}. \quad (11.11)$$

The interpolation error, which is $\mathcal{O}(h^{2k+2})$, may be neglected if $2k+2 > p+1$.

As to the choice of σ_i we denote the local error of the method by $le = y_1 - y(x_0 + h)$. It follows from Taylor expansion (see Theorem 3.2) that

$$y_{\sigma_i} - y(x_0 + \sigma_i h) = \sigma_i^{p+1} \cdot le + \mathcal{O}(h^{p+2})$$

and consequently the error of (11.10) satisfies (for $2k+2 > p+1$)

$$u(\theta) - y(x_0 + \theta h) = \left(\sum_{i=1}^k \sigma_i^{p+1} v_i(\theta) \right) \cdot le + \mathcal{O}(h^{p+2}). \quad (11.12)$$

The coefficient of le is equal to 1 for $\theta = 1$ and it is natural to search for suitable σ_i such that

$$\left| \sum_{i=1}^k \sigma_i^{p+1} v_i(\theta) \right| \leq 1 \quad \text{for all } \theta \in [0, 1]. \quad (11.13)$$

Indeed, under the assumption $2k-1 \leq p < 2k+1$, it can be shown that numbers $0 = \sigma_0 < \sigma_1 < \dots < \sigma_{k-1} < \sigma_k = 1$ exist satisfying (11.13) (see Exercise 1). Selected values of σ_i proposed by Enright & Higham (1991), which satisfy this condition are given in Table 11.1. For such a choice of σ_i the error (11.12) of the dense output is bounded (at least asymptotically) by the local error le at the endpoint of integration. This implementation of a dense output provides a simple way to estimate le . Since $u(\theta)$ is an $\mathcal{O}(h^{p+1})$ -approximation of $y(x_0 + \theta h)$, the defect of $u(\theta)$ satisfies

$$u'(\theta) - hf(x_0 + \theta h, u(\theta)) = \left(\sum_{i=1}^k \sigma_i^{p+1} v_i'(\theta) \right) \cdot le + \mathcal{O}(h^{p+2}). \quad (11.14)$$

If we take a σ^* different from σ_i such that $\sum_{i=1}^k \sigma_i^{p+1} v_i'(\sigma^*) \neq 0$ (see Table 11.1) then only one function evaluation, namely $f(x_0 + \sigma^* h, u(\sigma^*))$, allows the computation of an asymptotically correct approximation of le from (11.14). This error estimate can be used for step size selection and for improving the numerical result (local extrapolation). In the local extrapolation mode one then loses the \mathcal{C}^1 continuity of the dense output.

With the use of an additional processor the quantities y_{σ^*} and $f(x_0 + \sigma^* h, y_{\sigma^*})$ can be computed simultaneously with y_{σ_i} and $f(x_0 + \sigma_i h, y_{\sigma_i})$. If the polynomial $u(\theta)$ is required to satisfy $u(\sigma^*) = y_{\sigma^*}$, but not $u'(\sigma^*) = hf(x_0 + \sigma^* h, y_{\sigma^*})$, then the estimate (11.14) of the local error le does not need any further evaluation of f .

Table 11.1. Good values for σ_i

p	k	$\sigma_1, \dots, \sigma_{k-1}$	σ^*
5	3	0.2, 0.4	0.88
6	3	0.2, 0.4	0.88
7	4	0.2, 0.4, 0.7	0.94
8	4	0.2, 0.4, 0.6	0.93

Exercises

1. Let the positive integers k and p satisfy $2k - 1 \leq p < 2k + 1$. Then show that there exist numbers $0 = \sigma_0 < \sigma_1 < \dots < \sigma_{k-1} < \sigma_k = 1$ such that (11.13) is true for all $\theta \in [0, 1]$.

Hint. Put $\sigma_j = j\varepsilon$ for $j = 1, \dots, k - 1$ and show that (11.13) is verified for sufficiently small $\varepsilon > 0$. Of course, in a computer program, one should use σ_j which satisfy (11.13) and are well separated in order to avoid roundoff errors.

II.12 Composition of B-Series

At the Dundee Conference in 1969, a paper by J. Butcher was read which contained a surprising result. (H.J. Stetter 1971)

We shall now derive a theorem on the composition of what we call B-series (in honour of J. Butcher). This will have many applications and will lead to a better understanding of order conditions for all general classes of methods (composition of methods, multiderivative methods of Section II.13, general linear methods of Section III.8, Rosenbrock methods in Exercise 2 of Section IV.7).

Composition of Runge-Kutta Methods

There is no five-stage explicit Runge-Kutta method of order 5 (Section II.5). This led Butcher (1969) to the idea of searching for different five-stage methods such that a certain *composition* of these methods produces a fifth-order result (“effective order”). Although not of much practical interest (mainly due to the problem of changing step size), this was the starting point of a fascinating algebraic theory of numerical methods.

Suppose we have two methods, say of three stages,

$$\begin{array}{c|ccc}
 0 & & & \\
 \hat{c}_2 & \hat{a}_{21} & & \\
 \hat{c}_3 & \hat{a}_{31} & \hat{a}_{32} & \\
 \hline
 & \hat{b}_1 & \hat{b}_2 & \hat{b}_3
 \end{array}
 \qquad
 \begin{array}{c|ccc}
 0 & & & \\
 \tilde{c}_2 & \tilde{a}_{21} & & \\
 \tilde{c}_3 & \tilde{a}_{31} & \tilde{a}_{32} & \\
 \hline
 & \tilde{b}_1 & \tilde{b}_2 & \tilde{b}_3
 \end{array}
 \tag{12.1}$$

which are applied one after the other to a starting value y_0 with the same step size:

$$g_i = y_0 + h \sum_j \hat{a}_{ij} f(g_j), \quad y_1 = y_0 + h \sum_j \hat{b}_j f(g_j) \tag{12.2}$$

$$\ell_i = y_1 + h \sum_j \tilde{a}_{ij} f(\ell_j), \quad y_2 = y_1 + h \sum_j \tilde{b}_j f(\ell_j). \tag{12.3}$$

If we insert y_1 from (12.2) into (12.3) and group all g_i, ℓ_i together, we see that the

composition can be interpreted as a large Runge-Kutta method with coefficients

$$\begin{array}{c|cccc}
 0 & & & & \\
 \widehat{c}_2 & \widehat{a}_{21} & & & \\
 \widehat{c}_3 & \widehat{a}_{31} & \widehat{a}_{32} & & \\
 \sum \widehat{b}_i & \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_3 & \\
 \sum \widehat{b}_i + \widetilde{c}_2 & \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_3 & \widetilde{a}_{21} \\
 \sum \widehat{b}_i + \widetilde{c}_3 & \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_3 & \widetilde{a}_{31} \quad \widetilde{a}_{32} \\
 \hline
 & \widehat{b}_1 & \widehat{b}_2 & \widehat{b}_3 & \widetilde{b}_1 \quad \widetilde{b}_2 \quad \widetilde{b}_3
 \end{array}
 \equiv
 \begin{array}{c|cccccc}
 0 & & & & & \\
 c_2 & a_{21} & & & & \\
 c_3 & a_{31} & a_{32} & & & \\
 c_4 & a_{41} & a_{42} & a_{43} & & \\
 c_5 & a_{51} & a_{52} & a_{53} & a_{54} & \\
 c_6 & a_{61} & a_{62} & a_{63} & a_{64} & a_{65} \\
 \hline
 & b_1 & b_2 & b_3 & b_4 & b_5 \quad b_6
 \end{array}
 \quad (12.4)$$

It is now of interest to study the *order conditions* of the new method. For this, we have to compute the expressions (see Table 2.2)

$$\sum b_i, \quad 2 \sum b_i c_i, \quad 3 \sum b_i c_i^2, \quad 6 \sum b_i a_{ij} c_j, \quad \text{etc.}$$

If we insert the values from the left tableau of (12.4), a computation, which for low orders is still not too difficult, shows that these expressions can be written in terms of the corresponding expressions for the two methods (12.1). We shall denote these expressions for the *first* method by $\mathbf{a}(t)$, for the *second* method by $\mathbf{b}(t)$, and for the *composite* method by $\mathbf{ab}(t)$:

$$\mathbf{a}(\cdot) = \sum \widehat{b}_i, \quad \mathbf{a}(\cdot) = 2 \cdot \sum \widehat{b}_i \widehat{c}_i, \quad \mathbf{a}(\cdot) = 3 \cdot \sum \widehat{b}_i \widehat{c}_i^2, \quad \dots \quad (12.5a)$$

$$\mathbf{b}(\cdot) = \sum \widetilde{b}_i, \quad \mathbf{b}(\cdot) = 2 \cdot \sum \widetilde{b}_i \widetilde{c}_i, \quad \mathbf{b}(\cdot) = 3 \cdot \sum \widetilde{b}_i \widetilde{c}_i^2, \quad \dots \quad (12.5b)$$

$$\mathbf{ab}(\cdot) = \sum b_i, \quad \mathbf{ab}(\cdot) = 2 \cdot \sum b_i c_i, \quad \mathbf{ab}(\cdot) = 3 \cdot \sum b_i c_i^2, \quad \dots \quad (12.5c)$$

The above mentioned formulas are then

$$\begin{aligned}
 \mathbf{ab}(\cdot) &= \mathbf{a}(\cdot) + \mathbf{b}(\cdot) \\
 \mathbf{ab}(\cdot) &= \mathbf{a}(\cdot) + 2\mathbf{b}(\cdot)\mathbf{a}(\cdot) + \mathbf{b}(\cdot) \\
 \mathbf{ab}(\cdot) &= \mathbf{a}(\cdot) + 3\mathbf{b}(\cdot)\mathbf{a}(\cdot)^2 + 3\mathbf{b}(\cdot)\mathbf{a}(\cdot) + \mathbf{b}(\cdot) \\
 \mathbf{ab}(\cdot) &= \mathbf{a}(\cdot) + 3\mathbf{b}(\cdot)\mathbf{a}(\cdot) + 3\mathbf{b}(\cdot)\mathbf{a}(\cdot) + \mathbf{b}(\cdot)
 \end{aligned}
 \quad (12.6)$$

etc.

It is now, of course, of interest to have a general understanding of these formulas for arbitrary trees. This, however, is not easy in the above framework (“... a tedious calculation shows that ...”). Further, there are problems of identifying different methods with identical numerical results (see Exercise 1 below). Also, we want the theory to include more general processes than Runge-Kutta methods, for example the exact solution or multi-derivative methods.

B-Series

All these difficulties can be avoided if we consider directly the composition of the series appearing in Section II.2. We define by

$$T = \{\emptyset\} \cup T_1 \cup T_2 \cup \dots, \quad LT = \{\emptyset\} \cup LT_1 \cup LT_2 \cup \dots$$

the sets of all trees and labelled trees, respectively.

Definition 12.1 (Hairer & Wanner 1974). Let $\mathbf{a}(\emptyset)$, $\mathbf{a}(\cdot)$, $\mathbf{a}(\text{hook})$, $\mathbf{a}(\text{fishhook})$, \dots be a sequence of real coefficients defined for all trees $\mathbf{a} : T \rightarrow \mathbb{R}$. Then we call the series (see Theorem 2.11, Definitions 2.2, 2.3)

$$\begin{aligned} B(\mathbf{a}, y) &= \mathbf{a}(\emptyset)y + h\mathbf{a}(\cdot)f(y) + \frac{h^2}{2!}\mathbf{a}(\text{hook})F(\text{hook})(y) + \dots \\ &= \sum_{t \in LT} \frac{h^{\varrho(t)}}{\varrho(t)!} \mathbf{a}(t)F(t)(y) = \sum_{t \in T} \frac{h^{\varrho(t)}}{\varrho(t)!} \alpha(t) \mathbf{a}(t)F(t)(y) \end{aligned} \quad (12.7)$$

a *B-series*.

We have seen in Theorems 2.11 and 2.6 that the numerical solution of a Runge-Kutta method as well as the exact solution are B-series. The coefficients of the latter are all equal to 1.

Usually we are only interested in a finite number of terms of these series (only as high as the orders of the methods under consideration, or as far as f is differentiable) and all subsequent results are valid modulo error terms $\mathcal{O}(h^{k+1})$.

Definition 12.2. Let $t \in LT$ be a labelled tree of order $q = \varrho(t)$ and $0 \leq i \leq q$ be a fixed integer. Then we denote by $s_i(t) = s$ the *subtree* formed by the first i indices and by $d_i(t)$ (the *difference set*) the set of subtrees formed by the remaining indices. In the graphical representation we distinguish the subtree s by fat nodes and doubled lines.

Example 12.3. For the labelled tree $t = \text{hook}$ we have:

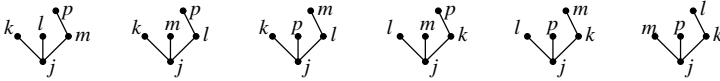
$i = 0$:		$s_0(t) = \emptyset,$	$d_0(t) = \{\text{fishhook}\}$
$i = 1$:		$s_1(t) = \cdot,$	$d_1(t) = \{\cdot, \cdot, \cdot, \text{hook}\}$
$i = 2$:		$s_2(t) = \text{hook},$	$d_2(t) = \{\cdot, \cdot, \cdot, \cdot\}$
$i = 3$:		$s_3(t) = \text{fishhook},$	$d_3(t) = \{\cdot, \cdot, \cdot\}$
$i = 4$:		$s_4(t) = \text{fishhook},$	$d_4(t) = \{\cdot\}$
$i = 5$:		$s_5(t) = t = \text{hook},$	$d_5(t) = \emptyset$

Definition 12.4. Let $\mathbf{a} : T \rightarrow \mathbb{R}$ and $\mathbf{b} : T \rightarrow \mathbb{R}$ be two sequences of coefficients such that $\mathbf{a}(\emptyset) = 1$. Then for a tree t of order $q = \varrho(t)$ we define the *composition*

$$\mathbf{ab}(t) = \frac{1}{\alpha(t)} \sum_{i=0}^q \left(\sum_i^q \binom{q}{i} \mathbf{b}(s_i(t)) \prod_{z \in d_i(t)} \mathbf{a}(z) \right) \quad (12.8)$$

where the first summation is over all $\alpha(t)$ different labellings of t (see Definition 2.5).

Example 12.5. It is easily seen that the formulas of (12.6) are special cases of (12.8). The tree t of Example 12.3 possesses 6 different labellings



These lead to

$$\begin{aligned} \mathbf{ab}(\heartsuit) &= \mathbf{b}(\emptyset)\mathbf{a}(\heartsuit) + 5\mathbf{b}(\cdot)\mathbf{a}(\cdot)^2\mathbf{a}(\cdot) \\ &\quad + 10\left(\frac{1}{2}\mathbf{b}(\cdot)\mathbf{a}(\cdot)\mathbf{a}(\cdot) + \frac{1}{2}\mathbf{b}(\cdot)\mathbf{a}(\cdot)^3\right) \\ &\quad + 10\left(\frac{1}{6}\mathbf{b}(\heartsuit)\mathbf{a}(\cdot) + \frac{4}{6}\mathbf{b}(\heartsuit)\mathbf{a}(\cdot)^2 + \frac{1}{6}\mathbf{b}(\heartsuit)\mathbf{a}(\cdot)^2\right) \\ &\quad + 5\left(\frac{1}{2}\mathbf{b}(\heartsuit)\mathbf{a}(\cdot) + \frac{1}{2}\mathbf{b}(\heartsuit)\mathbf{a}(\cdot)\right) + \mathbf{b}(\heartsuit). \end{aligned} \quad (12.9)$$

Here is the main theorem of this section:

Theorem 12.6 (Hairer & Wanner 1974). *As above, let $\mathbf{a} : T \rightarrow \mathbb{R}$ and $\mathbf{b} : T \rightarrow \mathbb{R}$ be two sequences of coefficients such that $\mathbf{a}(\emptyset) = 1$. Then the composition of the two corresponding B-series is again a B-series*

$$B(\mathbf{b}, B(\mathbf{a}, y)) = B(\mathbf{ab}, y) \quad (12.10)$$

where the “product” $\mathbf{ab} : T \rightarrow \mathbb{R}$ is that of Definition 12.4.

Proof. We denote the inner series by

$$B(\mathbf{a}, y) = g(h). \quad (12.11)$$

Then the proof is similar to the development of Section II.2 (see Fig. 2.2), with the difference that, instead of $f(g)$, we now start from

$$B(\mathbf{b}, g) = \sum_{s \in LT} \frac{h^{\varrho(s)}}{\varrho(s)!} \mathbf{b}(s) F(s)(g) \quad (12.12)$$

and have to compute the derivatives of this function: let us select the term $s = \heartsuit$

of this series,

$$\frac{h^3}{3!} \mathbf{b}(\mathfrak{z}) \sum_{L,M} f_L^K(g) f_M^L(g) f^M(g). \quad (12.13)$$

The q th derivative of this expression, for $h = 0$, is by Leibniz' formula

$$\binom{q}{3} \mathbf{b}(\mathfrak{z}) \sum_{L,M} (f_L^K(g) f_M^L(g) f^M(g))^{(q-3)}|_{h=0}. \quad (12.14)$$

We now compute, as we did in Lemma 2.8, the derivatives of

$$f_L^K(g) f_M^L(g) f^M(g) \quad (12.15)$$

using the classical rules of differential calculus; this gives for the first derivative

$$\sum_N f_{LN}^K \cdot (g^N)' f_M^L f^M + \sum_N f_L^K f_{MN}^L \cdot (g^N)' f^M + \sum_N f_L^K f_M^L f_N^M \cdot (g^N)'$$

and so on. We again represent this in graphical form in Fig. 12.1.

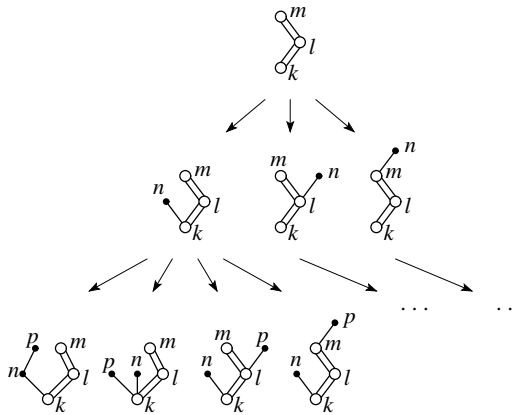


Fig. 12.1. Derivatives of (12.15)

We see that we arrive at trees u of order q such that $s_3(u) = s$ (where $3 = \varrho(s)$) and the elements of $d_3(u)$ have no ramifications. The corresponding expressions are similar to (2.6;q-1) in Lemma 2.8. We finally have to insert the derivatives of g (see (12.11)) and rearrange the terms. Then, as in Fig. 2.4, the tall branches of $d_3(u)$ are replaced by trees z of order δ , multiplied by $\mathbf{a}(z)$. Thus the coefficient which we obtain for a given tree t is just given by (12.8).

The factor $1/\alpha(t)$ is due to the fact that in $B(\mathbf{ab}, y)$ the term with $\mathbf{ab}(t)F(t)$ appears $\alpha(t)$ times. \square

Since $hf(y) = B(\mathbf{b}, y)$ is a special B-series with $\mathbf{b}(\cdot) = 1$ and all other $\mathbf{b}(t) = 0$, we have the following

Corollary 12.7. *If $\mathbf{a} : T \rightarrow \mathbb{R}$ with $\mathbf{a}(\emptyset) = 1$, then*

$$hf(B(\mathbf{a}, y)) = B(\mathbf{a}', y)$$

with

$$\begin{aligned} \mathbf{a}'(\emptyset) &= 0, \quad \mathbf{a}'(\cdot) = 1 \\ \mathbf{a}'([t_1, \dots, t_m]) &= \varrho(t) \mathbf{a}(t_1) \cdots \mathbf{a}(t_m) \end{aligned} \quad (12.16)$$

where $t = [t_1, \dots, t_m]$ means that $d_1(t) = \{t_1, t_2, \dots, t_m\}$ (Definition 2.12).

Proof. We obtain (12.16) from (12.8) with $i = 1$, $q = \varrho(t)$ and the fact that the expression in brackets is independent of the labelling of t . \square

Order Conditions for Runge-Kutta Methods

As an application of Corollary 12.7, we demonstrate the derivation of order conditions for Runge-Kutta methods: we write method (2.3) as

$$g_i = y_0 + \sum_{j=1}^s a_{ij} k_j, \quad k_i = hf(g_i), \quad y_1 = y_0 + \sum_{j=1}^s b_j k_j. \quad (12.17)$$

If we assume g_i , k_i and y_1 to be B-series, whose coefficients we denote by $\mathbf{g}_i, \mathbf{k}_i, \mathbf{y}_1$

$$g_i = B(\mathbf{g}_i, y_0), \quad k_i = B(\mathbf{k}_i, y_0), \quad y_1 = B(\mathbf{y}_1, y_0),$$

then Corollary 12.7 immediately allows us to transcribe formulas (12.17) as

$$\begin{aligned} \mathbf{g}_i(\emptyset) &= 1, & \mathbf{k}_i(\cdot) &= 1, & \mathbf{y}_1(\emptyset) &= 1, \\ \mathbf{g}_i(t) &= \sum_{j=1}^s a_{ij} \mathbf{k}_j(t), & \mathbf{k}_i(t) &= \varrho(t) \mathbf{g}_i(t_1) \cdots \mathbf{g}_i(t_m), & \mathbf{y}_1(t) &= \sum_{j=1}^s b_j \mathbf{k}_j(t) \end{aligned}$$

which leads easily to formulas (2.17), (2.19) and Theorem 2.11.

Also, if we put $y(h) = B(\mathbf{y}, y_0)$ for the *true solution*, and compare the derivative $hy'(h)$ of the series (12.7) with $hf(y(h))$ from Corollary 12.7, we immediately obtain $\mathbf{y}(t) = 1$ for all t , so that Theorem 2.6 drops out. The order conditions are then obtained as in Theorem 2.13 by comparing the coefficients of the B-series $B(\mathbf{y}, y_0)$ and $B(\mathbf{y}_1, y_0)$.

Butcher's "Effective Order"

We search for a 5-stage Runge-Kutta method **a** and for a method **d**, such that **dad**⁻¹ represents a fifth order method **u**. This means that we have to satisfy

$$\mathbf{da}(t) = \mathbf{yd}(t) \quad \text{for} \quad \varrho(t) \leq 5, \quad (12.18)$$

where **y**(*t*) = 1 represents the B-series of the exact solution. Then

$$(\mathbf{dad}^{-1})^k = \mathbf{da}^k \mathbf{d}^{-1} = (\mathbf{da}) \mathbf{a}^{k-2} (\mathbf{ad}^{-1}). \quad (12.19)$$

If now two Runge-Kutta methods **b** and **c** are constructed such that **b** = **da** and **c** = **ad**⁻¹ up to order 5, then applying one step of **b** followed by *k* - 2 steps of **a** and a final step of **c** is equivalent (up to order 5) to *k* steps of the 5th order method **dad**⁻¹ (see Fig. 12.2). A possible set of coefficients, computed by Butcher (1969), is given in Table 12.1 (method **a** has classical order 4).

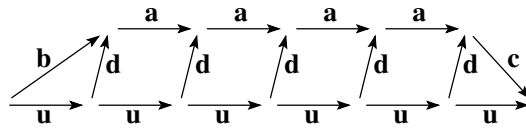


Fig. 12.2. Effective increase of order

Stetter's approach. Soon after the appearance of Butcher's purely algebraic proof, Stetter (1971) gave an elegant analytic explanation. Consider the principal global error term $e_p(x)$ which satisfies the variational equation (8.8). The question is, under which conditions on the local error $d_{p+1}(x)$ (see (8.8)) this equation can be solved, for special initial values, without effort. We write equation (8.8) as

$$e'(x) - \frac{\partial f}{\partial y}(y(x)) \cdot e(x) = d(x) \quad (12.20)$$

and want $e(x)$ to possess an expansion of the form

$$e(x) = \sum_{t \in T_p} \alpha(t) \mathbf{e}(t) F(t)(y(x)) \quad (12.21)$$

with constant coefficients $\mathbf{e}(t)$. Simply inserting (12.21) into (12.20) yields

$$d(x) = \sum_{t \in T_p} \alpha(t) \mathbf{e}(t) \left(\frac{d}{dx} (F(t)(y(x))) - f'(y(x)) \cdot F(t)(y(x)) \right). \quad (12.22)$$

Thus, (12.21) is the exact solution of the variational equation, if the local error $d(x)$ has the symmetric form (12.22). Then, if we replace the initial value y_0 by the "starting procedure"

$$\hat{y}_0 := y_0 - h^p e(x_0) = y_0 - h^p \sum_{t \in T_p} \alpha(t) \mathbf{e}(t) F(t)(y_0) \quad (12.23)$$

Table 12.1. Butcher's method of effective order 5

0	Method a				
$\frac{1}{5}$	$\frac{1}{5}$				
$\frac{2}{5}$	0	$\frac{2}{5}$			
$\frac{1}{2}$	$\frac{3}{16}$	0	$\frac{5}{16}$		
1	$\frac{1}{4}$	0	$-\frac{5}{4}$	2	
	$\frac{1}{6}$	0	0	$\frac{2}{3}$	$\frac{1}{6}$

0	Method b				0	Method c			
$\frac{1}{5}$	$\frac{1}{5}$				$\frac{1}{5}$	$\frac{1}{5}$			
$\frac{2}{5}$	0	$\frac{2}{5}$			$\frac{2}{5}$	0	$\frac{2}{5}$		
$\frac{3}{4}$	$\frac{75}{64}$	$-\frac{9}{4}$	$\frac{117}{64}$		$\frac{3}{4}$	$\frac{161}{192}$	$-\frac{19}{12}$	$\frac{287}{192}$	
1	$-\frac{37}{36}$	$\frac{7}{3}$	$-\frac{3}{4}$	$\frac{4}{9}$	1	$-\frac{27}{28}$	$\frac{19}{7}$	$-\frac{291}{196}$	$\frac{36}{49}$
	$\frac{19}{144}$	0	$\frac{25}{48}$	$\frac{2}{9}$	$\frac{1}{8}$	$\frac{7}{48}$	0	$\frac{475}{1008}$	$\frac{2}{7}$

(or by a Runge-Kutta method equivalent to this up to order $p+1$; this would represent “method **d**” in Fig. 12.2), its error satisfies $y(x_0) - \hat{y}_0 = h^p e(x_0) + \mathcal{O}(h^{p+1})$. By Theorem 8.1 the numerical solution \hat{y}_n of the Runge-Kutta method applied to \hat{y}_0 satisfies $y(x_n) - \hat{y}_n = h^p e(x_n) + \mathcal{O}(h^{p+1})$. Therefore the “finishing procedure”

$$y_n := \hat{y}_n + h^p e(x_n) = \hat{y}_n + h^p \sum_{t \in T_p} \alpha(t) \mathbf{e}(t) F(t)(\hat{y}_n) + \mathcal{O}(h^{p+1}) \quad (12.24)$$

(or some equivalent Runge-Kutta method) gives a $(p+1)$ th order approximation to the solution.

Example. Butcher's method **a** of Table 12.1 has the local error

$$d_6(x) = \frac{1}{6!} \left(-\frac{1}{24} F(\Psi) - \frac{1}{4} F(\Psi) - \frac{1}{8} F(\Psi) + \frac{1}{6} F(\Psi) + \frac{1}{2} F(\Psi) \right). \quad (12.25)$$

The right-hand side of (12.22) would be (the derivation $\frac{d}{dx} F$ attaches a new twig

to each of the nodes, the product $f'(y) \cdot F$ lifts the tree on a stilt)

$$\begin{aligned}
 & e(\Psi) \left(F(\Psi) + 3F(\Phi) - F(\Upsilon) \right) \\
 & + 3e(\Phi) \left(F(\Phi) + F(\Psi) + F(\Upsilon) + F(\dot{\Phi}) - F(\dot{\Upsilon}) \right) \\
 & + e(\Upsilon) \left(F(\Upsilon) + F(\Psi) + 2F(\dot{\Upsilon}) - F(\dot{\Psi}) \right) \\
 & + e(\dot{\Psi}) \left(F(\dot{\Psi}) + F(\dot{\Upsilon}) + F(\dot{\Phi}) + F(\dot{\dot{\Psi}}) - F(\dot{\dot{\Phi}}) \right).
 \end{aligned} \tag{12.26}$$

Comparison of (12.25) and (12.26) shows that this method does indeed have the desired symmetry if

$$e(\Psi) = e(\Phi) = -\frac{1}{6!} \cdot \frac{1}{24}, \quad e(\Upsilon) = e(\dot{\Psi}) = \frac{1}{6!} \cdot \frac{1}{8}.$$

This allows one to construct a Runge-Kutta method as starting procedure corresponding to (12.23) up to the desired order.

Exercises

1. Show that the pairs of methods given in Tables 12.2 - 12.4 produce, at least for h sufficiently small, identical numerical results.

Result. a) is seen by permutation of the stages, b) by neglecting superfluous stages (Dahlquist & Jeltsch 1979), c) by identifying equal stages (Stetter 1973, Hundsdorfer & Spijker 1981). See also the survey on “The Runge-Kutta space” by Butcher (1984).

2. Extend formulas (12.6) by computing the composition $\mathbf{ab}(t)$ for all trees of order 4 and 5.
3. Verify that the methods given in Table 12.1 satisfy the stated order properties.
4. Prove, using Theorem 12.6, that the set

$$G = \{\mathbf{a} : T \rightarrow \mathbb{R} \mid \mathbf{a}(\emptyset) = 1\}$$

together with the composition law of Definition 12.4 is a (non-commutative) group.

5. (Equivalence of Butcher’s and Stetter’s approach). Let $\mathbf{a} : T \rightarrow \mathbb{R}$ represent a Runge-Kutta method of classical order p and effective order $p+1$, i.e., $\mathbf{a}(t) = 1$ for $\varrho(t) \leq p$ and

$$\mathbf{da}(t) = \mathbf{yd}(t) \quad \text{for} \quad \varrho(t) \leq p+1 \tag{12.27}$$

for some $\mathbf{d} : T \rightarrow \mathbb{R}$ and with $\mathbf{y}(t)$ as in (12.18). Prove that then the local error $h^{p+1}d(x) + \mathcal{O}(h^{p+2})$ of the method \mathbf{a} has the symmetric form (12.22). This

Table 12.2. Equivalent methods a)

0				1	0	1
1	1	0		0	0	0
			1/4	3/4	3/4	1/4

Table 12.3. Equivalent methods b)

1	2	0	0	-1	1	2	-1
3	0	1	2	0	2	1	1
7	0	3	4	0			
2	1	0	0	1			
					1/2	0	1/2

Table 12.4. Equivalent methods c)

1	1	1	1	-2	1	3	-2
1	2	2	-1	-2	-1	2	-3
1	-1	-1	5	-2			
-1	-1	2	1	-3			
					3/4	1/4	

means that, in this situation, Butcher's effective order is equivalent to Stetter's approach.

Hint. Start by expanding condition (12.27) (using (12.8)) for the first trees. Possible simplifications are then best seen if the second sum $\sum_{i=0}^q$ (for $\mathbf{y}\mathbf{d}$) is arranged *downwards* ($i = q, q-1, \dots, 0$). One then arrives recursively at the result

$$\mathbf{d}(t) = \mathbf{d}(\cdot)^{\varrho(t)} \quad \text{for } \varrho(t) \leq p-1.$$

Then express the error coefficients $\mathbf{a}(t) - 1$ for $\varrho(t) = p+1$ in terms of $\mathbf{d}(s) - \mathbf{d}(\cdot)^{\varrho(s)}$ where $\varrho(s) = p$. Formula (12.22) then becomes visible.

6. Prove that for $t = [t_1, \dots, t_m]$ the coefficient $\alpha(t)$ of Definition 2.5 satisfies the recurrence relation

$$\alpha(t) = \binom{\varrho(t)-1}{\varrho(t_1), \dots, \varrho(t_m)} \alpha(t_1) \cdot \dots \cdot \alpha(t_m) \cdot \frac{1}{\mu_1! \mu_2! \dots}. \quad (12.28)$$

The integers μ_1, μ_2, \dots count the equal trees among t_1, \dots, t_m .

Hint. The multinomial coefficient in (12.28) counts the possible partitionings of the labels $2, \dots, \varrho(t)$ to the m subtrees t_1, \dots, t_m . Equal subtrees lead to equal labellings. Hence the division by $\mu_1! \mu_2! \dots$.

II.13 Higher Derivative Methods

In Section I.8 we studied the computation of higher derivatives of solutions of

$$(y^J)' = f^J(x, y^1, \dots, y^n), \quad J = 1, \dots, n. \quad (13.1)$$

The chain rule

$$(y^J)'' = \frac{\partial f^J}{\partial x}(x, y) + \frac{\partial f^J}{\partial y^1}(x, y) \cdot f^1(x, y) + \dots + \frac{\partial f^J}{\partial y^n}(x, y) \cdot f^n(x, y) \quad (13.2)$$

leads to the differential operator D which, when applied to a function $\Psi(x, y)$, is given by

$$(D\Psi)(x, y) = \frac{\partial \Psi}{\partial x}(x, y) + \frac{\partial \Psi}{\partial y^1}(x, y) \cdot f^1(x, y) + \dots + \frac{\partial \Psi}{\partial y^n}(x, y) \cdot f^n(x, y). \quad (13.2')$$

Since $Dy^J = f^J$, we see by extending (13.2) that

$$(y^J)^{(\ell)} = (D^\ell y^J)(x, y), \quad \ell = 0, 1, 2, \dots \quad (13.3)$$

This notation allows us to define a new class of methods which combine features of Runge-Kutta methods as well as Taylor series methods:

Definition 13.1. Let $a_{ij}^{(r)}$, $b_j^{(r)}$, ($i, j = 1, \dots, s$, $r = 1, \dots, q$) be real coefficients. Then the method

$$\begin{aligned} k_i^{(\ell)} &= \frac{h^\ell}{\ell!} (D^\ell y) \left(x_0 + c_i h, y_0 + \sum_{r=1}^q \sum_{j=1}^s a_{ij}^{(r)} k_j^{(r)} \right) \\ y_1 &= y_0 + \sum_{r=1}^q \sum_{j=1}^s b_j^{(r)} k_j^{(r)} \end{aligned} \quad (13.4)$$

is called an s -stage q -derivative Runge-Kutta method. If $a_{ij}^{(r)} = 0$ for $i \leq j$, the method is *explicit*, otherwise *implicit*.

A natural extension of (1.9) is here, because of $Dx = 1$, $D^\ell x = 0$ ($\ell \geq 2$),

$$c_i = \sum_{j=1}^s a_{ij}^{(1)}. \quad (13.5)$$

Definition 13.1 is from Kastlunger & Wanner (1972), but special methods of this type have been considered earlier in the literature. In particular, the very successful methods of Fehlberg (1958, 1964) have this structure.

Collocation Methods

A natural way of obtaining s -stage q -derivative methods is to use the collocation idea with *multiple nodes*, i.e., to replace (7.15b) by

$$u^{(\ell)}(x_0 + c_i h) = (D^\ell y)(x_0 + c_i h, u(x_0 + c_i h)) \quad i = 1, \dots, s, \quad \ell = 1, \dots, q_i \quad (13.6)$$

where $u(x)$ is a polynomial of degree $q_1 + q_2 + \dots + q_s$ and q_1, \dots, q_s , the “multiplicities” of the nodes c_1, \dots, c_s , are given integers. For example $q_1 = m$, $q_2 = \dots = q_s = 1$ leads to Fehlberg-type methods.

In order to generalize the results and ideas of Section II.7, we have to replace the Lagrange interpolation of Theorem 7.7 by *Hermite* interpolation (Hermite 1878: “Je me suis proposé de trouver un polynôme . . .”). The reason is that (13.6) can be interpreted as an ordinary collocation condition with clusters of q_i nodes “infinitely” close together (Rolle’s theorem). We write Hermite’s formula as

$$p(t) = \sum_{j=1}^s \sum_{r=1}^{q_j} \frac{1}{r!} \ell_{jr}(t) p^{(r-1)}(c_j) \quad (13.7)$$

for polynomials $p(t)$ of degree $\sum q_j - 1$. Here the “basis” polynomials $\ell_{jr}(t)$ of degree $\sum q_j - 1$ must satisfy

$$l_{jr}^{(k)}(c_i) = \begin{cases} r! & \text{if } i = j \text{ and } k = r - 1 \\ 0 & \text{else} \end{cases} \quad (13.8)$$

and are best obtained from Newton’s interpolation formula (with multiple nodes). We now use this formula, as we did in Section II.7, for $p(t) = hu'(x_0 + th)$:

$$hu'(x_0 + th) = \sum_{j=1}^s \sum_{r=1}^{q_j} \ell_{jr}(t) k_j^{(r)}, \quad (13.9)$$

with

$$k_j^{(r)} = \frac{h^r}{r!} u^{(r)}(x_0 + c_j h). \quad (13.10)$$

If we insert

$$u(x_0 + c_i h) = y_0 + \int_0^{c_i} hu'(x_0 + th) dt$$

together with (13.9) into (13.6), we get:

Theorem 13.2. *The collocation method (13.6) is equivalent to an s -stage q -derivative implicit Runge-Kutta method (13.4) with*

$$a_{ij}^{(r)} = \int_0^{c_i} \ell_{jr}(t) dt, \quad b_j^{(r)} = \int_0^1 \ell_{jr}(t) dt. \quad (13.11)$$

□

Theorems 7.8, 7.9, and 7.10 now generalize immediately to the case of “confluent” quadrature formulas; i.e., the q -derivative Runge-Kutta method possesses the *same order* as the underlying quadrature formula

$$\int_0^1 p(t) dt \approx \sum_{j=1}^s \sum_{r=1}^{q_j} b_j^{(r)} p^{(r-1)}(c_j).$$

The “algebraic” proof of this result (extending Exercise 7 of Section II.7) is more complicated and is given, for the case $q_j = q$, in Kastlunger & Wanner (1972b).

The formulas corresponding to condition $C(\eta)$ are given by

$$\sum_{j=1}^s \sum_{r=1}^{q_j} a_{ij}^{(r)} \binom{\varrho}{r} c_j^{\varrho-r} = c_i^{\varrho}, \quad \varrho = 1, 2, \dots, \sum_{j=1}^s q_j. \quad (13.12)$$

These equations uniquely determine the $a_{ij}^{(r)}$, once the c_i have been chosen, by a linear system with a “confluent” Vandermonde matrix (see e.g., Gautschi 1962). Formula (13.12) is obtained by setting $p(t) = t^{\varrho-1}$ in (13.7) and then integrating from 0 to c_i .

Examples of methods. “Gaussian” quadrature formulas with multiple nodes exist for *odd* q (Stroud & Stancu 1965) and extend to q -derivative implicit Runge-Kutta methods (Kastlunger & Wanner 1972b): for $s = 1$ we have, of course, $c_1 = 1/2$ which yields

$$b_1^{(2k)} = 0, \quad b_1^{(2k+1)} = 2^{-2k}, \quad a_{11}^{(k)} = (-1)^{k+1} 2^{-k}.$$

We give also the coefficients for the case $s = 2$ and $q_1 = q_2 = 3$. The nodes c_i and the weights $b_i^{(k)}$ are those of Stroud & Stancu. The method has order 8:

$$\begin{array}{ll} c_1 = 0.185394435825045 & c_2 = 1 - c_1 \\ b_1^{(1)} = 0.5 & b_2^{(1)} = b_1^{(1)} \\ b_1^{(2)}/2! = 0.0240729420844974 & b_2^{(2)} = -b_1^{(2)} \\ b_1^{(3)}/3! = 0.00366264960671727 & b_2^{(3)} = b_1^{(3)} \end{array}$$

$$\begin{aligned}
a_{ij}^{(1)} &= \begin{pmatrix} 0.201854115831005 & -0.0164596800059598 \\ 0.516459680005959 & 0.298145884168994 \end{pmatrix} \\
a_{ij}^{(2)} &= \begin{pmatrix} -0.0223466569080541 & 0.00868878773082417 \\ 0.0568346718998190 & -0.0704925410770490 \end{pmatrix} \\
a_{ij}^{(3)} &= \begin{pmatrix} 0.0116739668400997 & -0.00215351251065784 \\ 0.0241294101509615 & 0.0103019308002039 \end{pmatrix}
\end{aligned}$$

Hermite-Obreschkoff Methods

We now consider the special case of collocation methods with $s=2$, $c_1=0$, $c_2=1$. These methods can be obtained in closed form by repeated partial integration as follows (Darboux 1876, Hermite 1878):

Lemma 13.3. *Let m be a given positive integer and $P(t)$ a polynomial of exact degree m . Then*

$$\sum_{j=0}^m h^j (D^j y)(x_1, y_1) P^{(m-j)}(0) = \sum_{j=0}^m h^j (D^j y)(x_0, y_0) P^{(m-j)}(1) \quad (13.13)$$

defines a multiderivative method (13.4) of order m .

Proof. We let $y(x)$ be the exact solution and start from

$$h^{m+1} \int_0^1 y^{(m+1)}(x_0 + ht) P(1-t) dt = \mathcal{O}(h^{m+1}).$$

This integral is now transformed by repeated partial integration until all derivatives of the polynomial $P(1-t)$ are used up. This leads to

$$\sum_{j=0}^m h^j y^{(j)}(x_1) P^{(m-j)}(0) = \sum_{j=0}^m h^j y^{(j)}(x_0) P^{(m-j)}(1) + \mathcal{O}(h^{m+1}).$$

If this is subtracted from (13.13) we find the difference of the left-hand sides to be $\mathcal{O}(h^{m+1})$, which shows by the implicit function theorem that (13.13) determines y_1 to this order if $P^{(m)}$, which is a constant, is $\neq 0$. \square

The argument $1-t$ in P (instead of the more natural t) avoids the sign changes in the partial integrations.

A good choice for $P(t)$ is, of course, a polynomial for which most derivatives disappear at $t=0$ and $t=1$. Then the method (13.13) is, by keeping the same order m , most economical. We write

$$P(t) = \frac{t^k(t-1)^\ell}{(k+\ell)!}$$

and obtain

$$\begin{aligned}
 y_1 - \frac{\ell}{(k+\ell)} \frac{h}{1!} (Dy)(x_1, y_1) + \frac{\ell(\ell-1)}{(k+\ell)(k+\ell-1)} \frac{h^2}{2!} (D^2y)(x_1, y_1) \pm \dots \\
 = y_0 + \frac{k}{(k+\ell)} \frac{h}{1!} (Dy)(x_0, y_0) + \frac{k(k-1)}{(k+\ell)(k+\ell-1)} \frac{h^2}{2!} (D^2y)(x_0, y_0) + \dots
 \end{aligned} \tag{13.14}$$

which is a method of order $m = k + \ell$. After the ℓ th term in the first line and the k th term in the second line, the coefficients automatically become zero. Special cases of this method are:

$$\begin{aligned}
 k = 1, \quad \ell = 0 &: \text{explicit Euler} \\
 k \geq 1, \quad \ell = 0 &: \text{Taylor series} \\
 k = 0, \quad \ell = 1 &: \text{implicit Euler} \\
 k = 1, \quad \ell = 1 &: \text{trapezoidal rule.}
 \end{aligned}$$

Darboux and Hermite advocated the use of this formula for the approximations of functions, Obreschkoff (1940) for the computation of integrals, Loscalzo & Schoenberg (1967), Loscalzo (1969) as well as Nørsett (1974a) for the solution of differential equations.

Fehlberg Methods

Another class of multiderivative methods is due to Fehlberg (1958, 1964): the idea is to subtract from the solution of $y' = f(x, y)$, $y(x_0) = y_0$ m terms of the Taylor series (see Section I.8)

$$\hat{y}(x) := y(x) - \sum_{i=0}^m Y_i(x - x_0)^i, \tag{13.15}$$

and to solve the resulting differential equation $\hat{y}'(x) = \hat{f}(x, \hat{y}(x))$, where

$$\hat{f}(x, \hat{y}(x)) = f\left(x, \hat{y} + \sum_{i=0}^m Y_i(x - x_0)^i\right) - \sum_{i=1}^m Y_i i (x - x_0)^{i-1}, \tag{13.16}$$

by a Runge-Kutta method. Thus, knowing that the solution of (13.16) and its first m derivatives are zero at the initial value, we can achieve much higher orders.

In order to understand this, we develop the Taylor series of the solution for the non-autonomous case, as we did at the beginning of Section II.1. We thereby omit the hats and suppose the transformation (13.15) already carried out. We then have from (1.6) (see also Exercise 3 of Section II.2)

$$\begin{aligned}
 f &= 0, \\
 f_x + f_y f &= 0, \\
 f_{xx} + 2f_{xy}f + f_{yy}f^2 + f_y(f_x + f_y f) &= 0, \text{ etc.}
 \end{aligned}$$

These formulas recursively imply that $f = 0$, $f_x = 0$, \dots , $\partial^{m-1} f / \partial x^{m-1} = 0$. All elementary differentials of order $\leq m$ and most of those of higher orders then become zero and the corresponding order conditions can be omitted. The first non-zero terms are

$$\begin{aligned} & \frac{\partial^m f}{\partial x^m} && \text{for order } m+1, \\ & \frac{\partial^{m+1} f}{\partial x^{m+1}} \quad \text{and} \quad \frac{\partial f}{\partial y} \cdot \frac{\partial^m f}{\partial x^m} && \text{for order } m+2, \end{aligned}$$

and so on. The corresponding order conditions are then

$$\sum_{i=1}^s b_i c_i^m = \frac{1}{m+1}$$

for order $m+1$,

$$\sum_{i=1}^s b_i c_i^{m+1} = \frac{1}{m+2} \quad \text{and} \quad \sum_{i,j} b_i a_{ij} c_j^m = \frac{1}{(m+1)(m+2)}$$

for order $m+2$, and so on.

The condition $\sum a_{ij} = c_i$, which usually allows several terms of (1.6) to be grouped together, is not necessary, because all these other terms are zero.

A complete insight is obtained by considering the method as being *partitioned* applied to the *partitioned system* $y' = f(x, y)$, $x' = 1$. This will be explained in Section II.15 (see Fig. 15.3).

Example 13.4. A solution with $s = 3$ stages of the (seven) conditions for order $m+3$ is given by Fehlberg (1964). The choice $c_1 = c_3 = 1$ minimizes the numerical work for the evaluation of (13.16) and the other coefficients are then uniquely determined (see Table 13.1).

Fehlberg (1964) also derived an embedded method with two additional stages of orders $m+3$ ($m+4$). These methods were widely used in the sixties for scientific computations.

Table 13.1. Fehlberg, order $m+3$

1			$\theta = \frac{m+1}{m+3}$
θ	$\frac{\theta^m}{m+3}$		
1	$-\frac{1}{m+1}$	$\frac{2}{(m+1)\theta^m}$	
	0	$\frac{m+3}{2(m+1)(m+2)\theta^m}$	
			$\frac{1}{2(m+2)}$

General Theory of Order Conditions

For the same reason as in Section II.2 we assume that (13.1) is autonomous. The general form of the order conditions for method (13.4) was derived in the thesis of Kastlunger (see Kastlunger & Wanner 1972). It later became a simple application of the composition theorem for B-series (Hairer & Wanner 1974). The point is that from Theorem 2.6,

$$\frac{h^i}{i!} (D^i y)(y_0) = \sum_{t \in LT, \varrho(t)=i} \frac{h^i}{i!} F(t)(y_0) = B(\mathbf{y}^{(i)}, y_0) \quad (13.17)$$

is a B-series with coefficients

$$\mathbf{y}^{(i)}(t) = \begin{cases} 1 & \text{if } \varrho(t) = i \\ 0 & \text{otherwise.} \end{cases} \quad (13.18)$$

Thus, in extension of Corollary 12.7, we have

$$\frac{h^i}{i!} (D^i y)(B(\mathbf{a}, y_0)) = B(\mathbf{a}^{(i)}, y_0) \quad (13.19)$$

where, from formula (12.8) with $q = \varrho(t)$,

$$\mathbf{a}^{(i)}(t) = (\mathbf{a}\mathbf{y}^{(i)})(t) = \frac{1}{\alpha(t)} \binom{q}{i} \sum \prod_{z \in d_i(t)} \mathbf{a}(z), \quad (13.20)$$

and the sum is over all $\alpha(t)$ different labellings of t . This allows us to compute recursively the coefficients of the B-series which appear in (13.4).

Example 13.5. The tree $t = \spadesuit$ sketched in Fig. 13.1 possesses three different labellings, two of which produce the same difference set $d_2(t)$, so that formula (13.20) becomes

$$\mathbf{a}''(\spadesuit) = 2(2(\mathbf{a}(\cdot))^2 + \mathbf{a}(\spadesuit)). \quad (13.21)$$

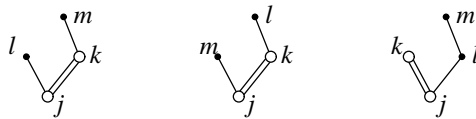


Fig. 13.1. Different labellings of \spadesuit

For all other trees of order ≤ 4 we have $\alpha(t) = 1$ and (13.20) leads to the following table of second derivatives

$$\begin{array}{ll} \mathbf{a}''(\cdot) = 0 & \mathbf{a}''(\spadesuit) = 1 \\ \mathbf{a}''(\heartsuit) = 3\mathbf{a}(\cdot) & \mathbf{a}''(\clubsuit) = 3\mathbf{a}(\cdot) \\ \mathbf{a}''(\blacklozenge) = 6(\mathbf{a}(\cdot))^2 & \mathbf{a}''(\spadesuit) = 4(\mathbf{a}(\cdot))^2 + 2\mathbf{a}(\spadesuit) \\ \mathbf{a}''(\heartsuit) = 6(\mathbf{a}(\cdot))^2 & \mathbf{a}''(\clubsuit) = 6\mathbf{a}(\spadesuit). \end{array} \quad (13.22)$$

Once these expressions have been established, we write formulas (13.4) in the form

$$k_i^{(\ell)} = \frac{h^\ell}{\ell!} (D^\ell y)(g_i)$$

$$g_i = y_0 + \sum_{r=1}^q \sum_{j=1}^s a_{ij}^{(r)} k_j^{(r)}, \quad y_1 = y_0 + \sum_{r=1}^q \sum_{j=1}^s b_j^{(r)} k_j^{(r)} \quad (13.23)$$

and suppose the expressions $k_i^{(\ell)}$, g_i , y_1 to be B-series

$$k_i^{(\ell)} = B(\mathbf{k}_i^{(\ell)}, y_0), \quad g_i = B(\mathbf{g}_i, y_0), \quad y_1 = B(\mathbf{y}_1, y_0).$$

Then equations (13.23) can be translated into

$$\mathbf{k}_i^{(1)}(t) = \varrho(t) \mathbf{g}_i(t_1) \cdot \dots \cdot \mathbf{g}_i(t_m), \quad \mathbf{k}_i^{(1)}(\tau) = 1 \quad (\text{see (12.16)})$$

$$\mathbf{k}_i^{(2)}(t) = \mathbf{g}_i''(t) \quad \text{from (13.22)}$$

$$\mathbf{k}_i^{(3)}(t) = \mathbf{g}_i'''(t) \quad \text{from Exercise 1 or Exercise 2, etc.}$$

$$\mathbf{g}_i(t) = \sum_{r=1}^q \sum_{j=1}^s a_{ij}^{(r)} \mathbf{k}_j^{(r)}(t), \quad \mathbf{y}_1(t) = \sum_{r=1}^q \sum_{j=1}^s b_j^{(r)} \mathbf{k}_j^{(r)}(t).$$

These formulas recursively determine all the coefficients. Method (13.4) (together with (13.5)) is then of order p if, as usual,

$$\mathbf{y}_1(t) = 1 \quad \text{for all } t \text{ with } \varrho(t) \leq p. \quad (13.24)$$

More details and special methods are given in Kastlunger & Wanner (1972); see also Exercise 3.

Exercises

1. Extend Example 13.5 and obtain formulas for $\mathbf{a}^{(3)}(t)$ for all trees of order ≤ 4 .
2. (Kastlunger). Prove the following variant form of formula (13.20) which extends (12.16) more directly and can also be used to obtain the formulas of Example 13.5. If $t = [t_1, \dots, t_m]$ then

$$\mathbf{a}^{(i)}(t) = \frac{\varrho(t)}{i} \sum_{\substack{\lambda_1 + \dots + \lambda_m = i-1 \\ \lambda_1, \dots, \lambda_m \geq 0}} \mathbf{a}^{(\lambda_1)}(t_1) \dots \mathbf{a}^{(\lambda_m)}(t_m)$$

Hint. See Kastlunger & Wanner (1972); Hairer & Wanner (1973), Section 5.

3. Show that the conditions for order 3 of method (13.4) are given by

$$\begin{aligned}\sum_i b_i^{(1)} &= 1 \\ 2 \sum_i b_i^{(1)} c_i + \sum_i b_i^{(2)} &= 1 \\ 3 \sum_i b_i^{(1)} c_i^2 + 3 \sum_i b_i^{(2)} c_i + \sum_i b_i^{(3)} &= 1 \\ 6 \sum_{i,j} b_i^{(1)} a_{ij}^{(1)} c_j + 3 \sum_i b_i^{(1)} e_i + 3 \sum_i b_i^{(2)} c_i + \sum_i b_i^{(3)} &= 1,\end{aligned}$$

where $c_i = \sum_j a_{ij}^{(1)}$, $e_i = \sum_j a_{ij}^{(2)}$.

4. (Zurmühl 1952, Albrecht 1955). Differentiate a given first order system of differential equations $y' = f(x, y)$ to obtain

$$y'' = (D^2 y)(x, y), \quad y(x_0) = y_0, \quad y'(x_0) = f_0.$$

Apply to this equation a special method for higher order systems (see the following Section II.14) to obtain higher-derivative methods. Show that the following method is of order six

$$\begin{aligned}k_1 &= h^2 g(x_0, y_0) \\ k_2 &= h^2 g\left(x_0 + \frac{h}{4}, y_0 + \frac{h}{4} f_0 + \frac{1}{32} k_1\right) \\ k_3 &= h^2 g\left(x_0 + \frac{h}{2}, y_0 + \frac{h}{2} f_0 + \frac{1}{24} (-k_1 + 4k_2)\right) \\ k_4 &= h^2 g\left(x_0 + \frac{3h}{4}, y_0 + \frac{3h}{4} f_0 + \frac{1}{32} (3k_1 + 4k_2 + 2k_3)\right) \\ y_1 &= y_0 + h f_0 + \frac{1}{90} (7k_1 + 24k_2 + 6k_3 + 8k_4)\end{aligned}$$

where $g(x, y) = (D^2 y)(x, y) = Df(x, y) = f_x(x, y) + f_y(x, y) \cdot f(x, y)$.

II.14 Numerical Methods for Second Order Differential Equations

Mutationem motus proportionalem esse vi motrici impressae
(Newton's Lex II, 1687)

Many differential equations which appear in practice are systems of the *second order*

$$y'' = f(x, y, y'). \quad (14.1)$$

This is mainly due to the fact that the forces are proportional to acceleration, i.e., to second derivatives. As mentioned in Section I.1, such a system can be transformed into a first order differential equation of doubled dimension by considering the vector (y, y') as the new variable:

$$\begin{pmatrix} y \\ y' \end{pmatrix}' = \begin{pmatrix} y' \\ f(x, y, y') \end{pmatrix} \quad \begin{matrix} y(x_0) = y_0 \\ y'(x_0) = y'_0 \end{matrix} \quad (14.2)$$

In order to solve (14.1) numerically, one can for instance apply a Runge-Kutta method (explicit or implicit) to (14.2). This yields

$$\begin{aligned} k_i &= y'_0 + h \sum_{j=1}^s a_{ij} k'_j \\ k'_i &= f(x_0 + c_i h, y_0 + h \sum_{j=1}^s a_{ij} k_j, y'_0 + h \sum_{j=1}^s a_{ij} k'_j) \\ y_1 &= y_0 + h \sum_{i=1}^s b_i k_i, \quad y'_1 = y'_0 + h \sum_{i=1}^s b_i k'_i. \end{aligned} \quad (14.3)$$

If we insert the first formula of (14.3) into the others we obtain (assuming (1.9) and an order ≥ 1)

$$\begin{aligned} k'_i &= f(x_0 + c_i h, y_0 + c_i h y'_0 + h^2 \sum_{j=1}^s \bar{a}_{ij} k'_j, y'_0 + h \sum_{j=1}^s a_{ij} k'_j) \\ y_1 &= y_0 + h y'_0 + h^2 \sum_{i=1}^s \bar{b}_i k'_i, \quad y'_1 = y'_0 + h \sum_{i=1}^s b_i k'_i \end{aligned} \quad (14.4)$$

where

$$\bar{a}_{ij} = \sum_{k=1}^s a_{ik} a_{kj}, \quad \bar{b}_i = \sum_{j=1}^s b_j a_{ji}. \quad (14.5)$$

For an implementation the representation (14.4) is preferable to (14.3), since about half of the storage can be saved. This may be important, in particular if the dimension of equation (14.1) is large.

Nyström Methods

R.H. Merson: "... I have not seen the paper by Nyström. Was it in English?"

J.M. Bennett: "In German actually, not Finnish."

(From the discussion following a talk of Merson 1957)

E.J. Nyström (1925) was the first to consider methods of the form (14.4) in which the coefficients do not necessarily satisfy (14.5) ("Da bis jetzt die *direkte* Anwendung der Rungeschen Methode auf den wichtigen Fall von Differentialgleichungen zweiter Ordnung nicht behandelt war ...". Nyström, 1925). Such direct methods are called *Nyström methods*.

Definition 14.1. A Nyström method (14.4) has *order* p if for sufficiently smooth problems (14.1)

$$y(x_0 + h) - y_1 = \mathcal{O}(h^{p+1}), \quad y'(x_0 + h) - y'_1 = \mathcal{O}(h^{p+1}). \quad (14.6)$$

An example of an explicit Nyström method where condition (14.5) is violated is given in Table 14.1. Nyström claimed that this method would be simpler to apply than "Runge-Kutta's" and reduce the work by about 25%. This is, of course, not true if the Runge-Kutta method is applied as in (14.4) (see also Exercise 2).

Table 14.1. Nyström, order 4

c_i	0								
	$\frac{1}{2}$	$\frac{1}{8}$	\bar{a}_{ij}			$\frac{1}{2}$	a_{ij}		
	$\frac{1}{2}$	$\frac{1}{8}$	0			0	$\frac{1}{2}$		
	1	0	0	$\frac{1}{2}$		0	0	1	
$\bar{b}_i \rightarrow$		$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	0	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6} \leftarrow b_i$

A *real* improvement can be achieved in the case where the right-hand side of (14.1) does not depend on y' , i.e.,

$$y'' = f(x, y). \quad (14.7)$$

Here the Nyström method becomes

$$\begin{aligned}
 k'_i &= f(x_0 + c_i h, y_0 + c_i h y'_0 + h^2 \sum_{j=1}^s \bar{a}_{ij} k'_j) \\
 y_1 &= y_0 + h y'_0 + h^2 \sum_{i=1}^s \bar{b}_i k'_i, \quad y'_1 = y'_0 + h \sum_{i=1}^s b_i k'_i,
 \end{aligned} \tag{14.8}$$

and the coefficients a_{ij} are no longer needed. Some examples are given in Table 14.2. The fifth-order method of Table 14.2 needs only four evaluations of f . This is a considerable improvement compared to Runge-Kutta methods where at least six evaluations are necessary (cf. Theorem 5.1).

Table 14.2. Methods for $y'' = f(x, y)$

Nyström, order 4				Nyström, order 5						
	0	\bar{a}_{ij}			0					
	$\frac{1}{2}$	$\frac{1}{8}$				$\frac{1}{50}$	\bar{a}_{ij}			
c_i	$\frac{1}{2}$	$\frac{1}{8}$				$\frac{2}{3}$	$\frac{-1}{27}$	$\frac{7}{27}$		
	1	0	$\frac{1}{2}$			1	$\frac{3}{10}$	$\frac{-2}{35}$	$\frac{9}{35}$	
	\bar{b}_i	$\frac{1}{6}$	$\frac{1}{3}$	0		\bar{b}_i	$\frac{14}{336}$	$\frac{100}{336}$	$\frac{54}{336}$	0
	b_i	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$		b_i	$\frac{14}{336}$	$\frac{125}{336}$	$\frac{162}{336}$	$\frac{35}{336}$

Global convergence. Introducing the variable $z_n = (y_n, y'_n)^T$, a Nyström method (14.4) can be written in the form

$$z_1 = z_0 + h\Phi(x_0, z_0, h) \tag{14.9}$$

where

$$\Phi(x_0, z_0, h) = \begin{pmatrix} y'_0 + h \sum_i \bar{b}_i k'_i \\ \sum_i b_i k'_i \end{pmatrix}.$$

(14.9) is just a special one-step method for the differential equation (14.2). For a p th order Nyström method the local error $(y(x_0 + h) - y_1, y'(x_0 + h) - y'_1)^T$ can be bounded by Ch^{p+1} (Definition 14.1), which is in agreement with formula (3.27). The convergence theorems of Section II.3 and the results on asymptotic expansions of the global error (Section II.8) are also valid here.

Our next aim is to derive the order conditions for Nyström methods. For this purpose we extend the theory of Section II.2 to second order differential equations (Hairer & Wanner 1976).

The Derivatives of the Exact Solution

As for first order equations we may restrict ourselves to systems of autonomous differential equations

$$(y^J)'' = f^J(y^1, \dots, y^n, y'^1, \dots, y'^n) \quad (14.10)$$

(if necessary, add $x'' = 0$). The superscript index J denotes the J th component of the corresponding vector. We now calculate the derivatives of the exact solution of (14.10). The second derivative is given by (14.10):

$$(y^J)^{(2)} = f^J(y, y'). \quad (14.11;2)$$

A repeated differentiation of this equation, using (14.10), leads to

$$(y^J)^{(3)} = \sum_K \frac{\partial f^J}{\partial y^K} (y, y') \cdot y'^K + \sum_K \frac{\partial f^J}{\partial y'^K} (y, y') f^K(y, y') \quad (14.11;3)$$

$$\begin{aligned} (y^J)^{(4)} = & \sum_{K,L} \frac{\partial^2 f^J}{\partial y^K \partial y^L} (y, y') \cdot y'^K \cdot y'^L & (14.11;4) \\ & + \sum_{K,L} \frac{\partial^2 f^J}{\partial y^K \partial y'^L} (y, y') \cdot y'^K \cdot f^L(y, y') + \sum_K \frac{\partial f^J}{\partial y^K} (y, y') f^K(y, y') \\ & + \sum_{K,L} \frac{\partial^2 f^J}{\partial y'^K \partial y^L} (y, y') f^K(y, y') \cdot y'^L \\ & + \sum_{K,L} \frac{\partial^2 f^J}{\partial y'^K \partial y'^L} (y, y') f^K(y, y') f^L(y, y') \\ & + \sum_{K,L} \frac{\partial f^J}{\partial y'^K} (y, y') \frac{\partial f^K}{\partial y^L} (y, y') y'^L \\ & + \sum_{K,L} \frac{\partial f^J}{\partial y^K} (y, y') \frac{\partial f^K}{\partial y'^L} (y, y') f^L(y, y') \end{aligned}$$

The continuation of this process becomes even more complex than for first order differential equations. A graphical representation of the above formulas will therefore be very helpful. In order to distinguish the derivatives with respect to y and y' we need two kinds of vertices: “meagre” and “fat”. Fig. 14.1 shows the graphs that correspond to the above formulas.

Definition 14.2. A labelled N -tree of order q is a labelled tree (see Definition 2.2)

$$t : A_q \setminus \{j\} \rightarrow A_q$$

together with a mapping

$$t' : A_q \rightarrow \{\text{“meagre”}, \text{“fat”}\}$$

$$\circ_j \quad (14.11;2)$$

$$\begin{array}{c} k \\ \bullet \\ \circ_j \end{array} \quad \begin{array}{c} k \\ \circ \\ \circ_j \end{array} \quad (14.11;3)$$

$$\begin{array}{c} k \\ \bullet \\ \circ_j \end{array} \begin{array}{c} l \\ \bullet \\ \circ_j \end{array} \quad \begin{array}{c} k \\ \bullet \\ \circ_j \end{array} \begin{array}{c} l \\ \circ \\ \circ_j \end{array} \quad \begin{array}{c} k \\ \bullet \\ \circ_j \end{array} \begin{array}{c} l \\ \circ \\ \circ_j \end{array} \quad \begin{array}{c} k \\ \circ \\ \circ_j \end{array} \begin{array}{c} l \\ \bullet \\ \circ_j \end{array} \quad \begin{array}{c} k \\ \circ \\ \circ_j \end{array} \begin{array}{c} l \\ \circ \\ \circ_j \end{array} \quad \begin{array}{c} k \\ \circ \\ \circ_j \end{array} \begin{array}{c} l \\ \bullet \\ \circ_j \end{array} \begin{array}{c} l \\ \circ \\ \circ_j \end{array} \quad (14.11;4)$$

$$\begin{array}{c} k \\ \bullet \\ \circ_j \end{array} \begin{array}{c} l \\ \bullet \\ \circ_j \end{array} \begin{array}{c} m \\ \bullet \\ \circ_j \end{array} \quad \begin{array}{c} k \\ \bullet \\ \circ_j \end{array} \begin{array}{c} l \\ \bullet \\ \circ_j \end{array} \begin{array}{c} m \\ \circ \\ \circ_j \end{array} \quad \begin{array}{c} k \\ \bullet \\ \circ_j \end{array} \begin{array}{c} l \\ \circ \\ \circ_j \end{array} \begin{array}{c} m \\ \bullet \\ \circ_j \end{array} \quad \begin{array}{c} k \\ \circ \\ \circ_j \end{array} \begin{array}{c} l \\ \bullet \\ \circ_j \end{array} \begin{array}{c} m \\ \bullet \\ \circ_j \end{array} \quad \begin{array}{c} k \\ \circ \\ \circ_j \end{array} \begin{array}{c} l \\ \bullet \\ \circ_j \end{array} \begin{array}{c} m \\ \circ \\ \circ_j \end{array} \quad \dots \quad (14.11;5)$$

Fig. 14.1. The derivatives of the exact solution

which satisfies:

a) the root of t is always fat; i.e., $t'(j) = \text{“fat”}$;

b) a meagre vertex has at most one son and this son has to be fat.

We denote by LNT_q the set of all labelled N-trees of order q .

The reason for condition (b) is that all derivatives of $g(y, y') = y'$ vanish identically with the exception of the first derivative with respect to y' .

In the sequel we use the notation *end-vertex* for a vertex which has no son. If no confusion is possible, we write t instead of (t, t') for a labelled N-tree.

Definition 14.3. For a labelled N-tree t we denote by

$$F^J(t)(y, y')$$

the expression which is a *sum* over the indices of all fat vertices of t (without “ j ”, the index of the root) and over the indices of all meagre end-vertices. The *general term* of this sum is a product of expressions

$$\frac{\partial^r f^K}{\partial y^L \dots \partial y'^M \dots} (y, y') \quad \text{and} \quad y'^K. \quad (14.12)$$

A factor of the first type appears if the fat vertex k is connected via a meagre son with l, \dots and directly with a fat son m, \dots ; a factor y'^K appears if “ k ” is the index of a meagre end-vertex. The vector $F(t)(y, y')$ is again called an *elementary differential*.

For some examples see Table 14.3 below. Observe that the indices of the meagre vertices, which are not end-vertices, play no role in the above definition. In analogy to Definition 2.4 we have

Definition 14.4. Two labelled N-trees (t, t') and (u, u') are *equivalent*, if they differ only by a permutation of their indices; i.e., if they have the same order, say

q , and if there exists a bijection $\sigma : A_q \rightarrow A_q$ with $\sigma(j) = j$, such that $t\sigma = \sigma u$ on $A_q \setminus \{j\}$ and $t'\sigma = u'$.

For example, the second and fourth labelled N-trees of formula (14.11;4) in Fig. 14.1 are equivalent; and also the second and fifth of formula (14.11;5).

Definition 14.5. An equivalence class of q th order labelled N-trees is called an *N-tree of order q* . The set of all N-trees of order q is denoted by NT_q . We further denote by $\alpha(t)$ the number of elements in the equivalence class t , i.e., the number of possible different monotonic labellings of t .

Representatives of N-trees up to order 5 are shown in Table 14.3. We are now able to give a closed formula for the derivatives of the exact solution of (14.10).

Theorem 14.6. *The exact solution of (14.10) satisfies*

$$y^{(q)} = \sum_{t \in LNT_{q-1}} F(t)(y, y') = \sum_{t \in NT_{q-1}} \alpha(t) F(t)(y, y'). \quad (14.11;q)$$

Proof. The general formula is obtained by continuing the computation for (14.11;2-4) as in Section II.2. \square

The Derivatives of the Numerical Solution

We first rewrite (14.4) as

$$\begin{aligned} g_i &= y_0 + c_i h y'_0 + \sum_{j=1}^s \bar{a}_{ij} h^2 f(g_j, g'_j), & g'_i &= y'_0 + \sum_{j=1}^s a_{ij} h f(g_j, g'_j) \\ y_1 &= y_0 + h y'_0 + \sum_{i=1}^s \bar{b}_i h^2 f(g_i, g'_i), & y'_1 &= y'_0 + \sum_{i=1}^s b_i h f(g_i, g'_i) \end{aligned} \quad (14.13)$$

so that the intermediate values g_i, g'_i are treated in the same way as y_1, y'_1 . In (14.13) there appear expressions of the form $h^2 \varphi(h)$ and $h \varphi(h)$. Therefore we have to use in addition to (2.4) the formula

$$(h^2 \varphi(h))^{(q)} \big|_{h=0} = q \cdot (q-1) \cdot (\varphi(h))^{(q-2)} \big|_{h=0}. \quad (14.14)$$

We now compute successively the derivatives of g_i^J and $g_i'^J$ at $h=0$:

$$(g_i^J)^{(1)} \big|_{h=0} = c_i y_0'^J \quad (14.15;1)$$

$$(g_i'^J)^{(1)} \big|_{h=0} = \sum_j a_{ij} f^J \big|_{y_0, y_0'} \quad (14.16;1)$$

$$(g_i^J)^{(2)}|_{h=0} = 2 \sum_j \bar{a}_{ij} f^J|_{y_0, y'_0}. \quad (14.15;2)$$

For a further differentiation we need

$$(f^J(g_j, g'_j))^{(1)} = \sum_K \frac{\partial f^J}{\partial y^K} (g_j, g'_j) (g_j^K)^{(1)} + \sum_K \frac{\partial f^J}{\partial y'^K} (g_j, g'_j) (g_j'^K)^{(1)}. \quad (14.17)$$

With this formula we then obtain

$$\begin{aligned} (g_i'^J)^{(2)}|_{h=0} &= 2 \sum_j a_{ij} c_j \sum_K \frac{\partial f^J}{\partial y^K} \cdot y'^K|_{y_0, y'_0} \\ &\quad + 2 \sum_{j,k} a_{ij} a_{jk} \sum_K \frac{\partial f^J}{\partial y'^K} \cdot f^K|_{y_0, y'_0} \end{aligned} \quad (14.16;2)$$

$$\begin{aligned} (g_i^J)^{(3)}|_{h=0} &= 3 \cdot 2 \sum_j \bar{a}_{ij} c_j \sum_K \frac{\partial f^J}{\partial y^K} \cdot y'^K|_{y_0, y'_0} \\ &\quad + 3 \cdot 2 \sum_{j,k} \bar{a}_{ij} a_{jk} \sum_K \frac{\partial f^J}{\partial y'^K} \cdot f^K|_{y_0, y'_0}. \end{aligned} \quad (14.15;3)$$

To write down a general formula we need

Definition 14.7. For a labelled N-tree we denote by $\Phi_j(t)$ the expression which is a sum over the indices of all fat vertices of t (without “ j ”, the index of the root). The general term of the sum is a product of

- a_{kl} if the fat vertex “ k ” has a fat son “ l ”;
- \bar{a}_{kl} if the fat vertex “ k ” is connected via a meagre son with “ l ”; and
- c_k^m if the fat vertex “ k ” is connected with m meagre end-vertices.

Theorem 14.8. The g_i, g'_i of (14.13) satisfy

$$(g_i)^{(q+1)}|_{h=0} = (q+1) \sum_{t \in LNT_q} \gamma(t) \sum_{j=1}^s \bar{a}_{ij} \Phi_j(t) F(t)(y_0, y'_0) \quad (14.15;q+1)$$

$$(g'_i)^{(q)}|_{h=0} = \sum_{t \in LNT_q} \gamma(t) \sum_{j=1}^s a_{ij} \Phi_j(t) F(t)(y_0, y'_0) \quad (14.16;q)$$

where $\gamma(t)$ is given in Definition 2.10.

Proof. For small values of q these formulas were obtained above; for general values of q they are proved like Theorem 2.11. System (14.2) is a special case of what will later be treated as a *partitioned system* (see Section II.15). Theorem 14.8 will then appear again in a new light. \square

Because of the similarity of the formulas for g_i and y_1 , g'_i and y'_1 we have

Theorem 14.9. *The numerical solution y_1, y'_1 of (14.13) satisfies*

$$(y_1)^{(q)}|_{h=0} = q \sum_{t \in LNT_{q-1}} \gamma(t) \sum_{i=1}^s \bar{b}_i \Phi_i(t) F(t)(y_0, y'_0) \quad (14.18; q)$$

$$(y'_1)^{(q-1)}|_{h=0} = \sum_{t \in LNT_{q-1}} \gamma(t) \sum_{i=1}^s b_i \Phi_i(t) F(t)(y_0, y'_0). \quad (14.19; q-1)$$

□

The Order Conditions

For the study of the order of a Nyström method (Definition 14.1) one has to compare the Taylor series of y_1, y'_1 with that of the true solution $y(x_0 + h), y'(x_0 + h)$.

Theorem 14.10. *A Nyström method (14.4) is of order p iff*

$$\sum_{i=1}^s \bar{b}_i \Phi_i(t) = \frac{1}{(\varrho(t) + 1) \cdot \gamma(t)} \quad \text{for } N\text{-trees } t \text{ with } \varrho(t) \leq p-1, \quad (14.20)$$

$$\sum_{i=1}^s b_i \Phi_i(t) = \frac{1}{\gamma(t)} \quad \text{for } N\text{-trees } t \text{ with } \varrho(t) \leq p. \quad (14.21)$$

Here $\varrho(t)$ denotes the order of the N -tree t , $\Phi_i(t)$ and $\gamma(t)$ are given by Definition 14.7 and formula (2.17).

Proof. The “if” part is an immediate consequence of Theorems 14.6 and 14.9. The “only if” part can be shown in the same way as for first order equations (cf. Exercise 4 of Section II.2). □

Let us briefly discuss whether the extra freedom in the choice of the parameters of (14.4) (by discarding the assumption (14.5)) can lead to a considerable improvement. Since the order conditions for Runge-Kutta methods (Theorem 2.13) are a subset of (14.21) (see Exercise 3 below), it is impossible to gain order with this extra freedom. Only some (never all) error coefficients can be made smaller. Therefore we shall turn to Nyström methods (14.8) for special second order differential equations (14.7).

For the study of the order conditions for (14.8) we write (14.7) in autonomous form

$$y'' = f(y). \quad (14.22)$$

This special form implies that those elementary differentials which contain derivatives with respect to y' vanish identically. Consequently, only the following subset of N-trees has to be considered:

Definition 14.11. An N-tree t is called a *special N-tree* or *SN-tree*, if the fat vertices have only meagre sons.

Theorem 14.12. A Nyström method (14.8) for the special differential equation (14.7) is of order p , iff

$$\sum_{i=1}^s \bar{b}_i \Phi_i(t) = \frac{1}{(\varrho(t) + 1) \cdot \gamma(t)} \quad \text{for SN-trees } t \text{ with } \varrho(t) \leq p - 1, \quad (14.23)$$

$$\sum_{i=1}^s b_i \Phi_i(t) = \frac{1}{\gamma(t)} \quad \text{for SN-trees } t \text{ with } \varrho(t) \leq p. \quad (14.24) \quad \square$$

All SN-trees up to order 5, together with the elementary differentials and the expressions Φ_j , ϱ , α , and γ , which are needed for the order conditions, are given in Table 14.3.

Higher order systems. The extension of the ideas of this section to *higher order* systems

$$y^{(n)} = f(x, y, y', \dots, y^{(n-1)}) \quad (14.25)$$

is now more or less straightforward. Again, a real improvement is only possible in the case when the right-hand side of (14.25) depends only on x and y . A famous paper on this subject is the work of Zurmühl (1948). Tables of order conditions and methods are given in Hebsacker (1982).

On the Construction of Nyström Methods

The following simplifying assumptions are useful for the construction of Nyström methods.

Lemma 14.13. Under the assumption

$$\bar{b}_i = b_i(1 - c_i) \quad i = 1, \dots, s \quad (14.26)$$

the condition (14.24) implies (14.23).

Proof. Let t be an SN-tree of order $\leq p - 1$ and denote by u the SN-tree of order $\varrho(t) + 1$ obtained from t by attaching a new branch with a meagre vertex to the root of t . By Definition 14.7 we have $\Phi_i(u) = c_i \Phi_i(t)$ and from formula (2.17) it

Table 14.3. SN-trees, elementary differentials and order conditions

t	graph	$\varrho(t)$	$\alpha(t)$	$\gamma(t)$	$F^J(t)(y, y')$	$\Phi_j(t)$
t_1		1	1	1	f^J	1
t_2		2	1	2	$\sum_K f_K^J y'^K$	c_j
t_3		3	1	3	$\sum_{K,L} f_{KL}^J y'^K y'^L$	c_j^2
t_4		3	1	6	$\sum_L f_L^J f^L$	$\sum_l \bar{a}_{jl}$
t_5		4	1	4	$\sum_{K,L,M} f_{KLM}^J y'^K y'^L y'^M$	c_j^3
t_6		4	3	8	$\sum_{L,M} f_{LM}^J y'^L f^M$	$\sum_m c_j \bar{a}_{jm}$
t_7		4	1	24	$\sum_{L,M} f_L^J f_{LM}^J y'^M$	$\sum_l \bar{a}_{jl} c_l$
t_8		5	1	5	$\sum_{K,L,M,P} f_{KLMP}^J y'^K y'^L y'^M y'^P$	c_j^4
t_9		5	6	10	$\sum_{L,M,P} f_{LMP}^J y'^L y'^M f^P$	$\sum_p c_j^2 \bar{a}_{jp}$
t_{10}		5	3	20	$\sum_{M,P} f_{MP}^J f^M f^P$	$\sum_{m,p} \bar{a}_{jm} \bar{a}_{jp}$
t_{11}		5	4	30	$\sum_{L,M,P} f_{LP}^J f_{LM}^J y'^M y'^P$	$\sum_l c_j \bar{a}_{jl} c_l$
t_{12}		5	1	60	$\sum_{L,M,P} f_L^J f_{MP}^J y'^M y'^P$	$\sum_l \bar{a}_{jl} c_l^2$
t_{13}		5	1	120	$\sum_{L,P} f_L^J f_P^L f^P$	$\sum_{l,p} \bar{a}_{jl} \bar{a}_{lp}$

follows that $\gamma(u) = (\varrho(t) + 1)\gamma(t)/\varrho(t)$. The conclusion now follows since

$$\sum_{i=1}^s \bar{b}_i \Phi_i(t) = \sum_{i=1}^s b_i \Phi_i(t) - \sum_{i=1}^s b_i \Phi_i(u) = \frac{1}{\gamma(t)} - \frac{1}{\gamma(u)} = \frac{1}{(\varrho(t) + 1)\gamma(t)}.$$

□

Lemma 14.14. *Let t and u be two SN-trees as sketched in Fig. 14.2, where the encircled parts are assumed to be identical. Then under the assumption*

$$\sum_{j=1}^s \bar{a}_{ij} = \frac{c_i^2}{2} \quad i = 1, \dots, s \quad (14.27)$$

the order conditions for t and u are the same.

Proof. It follows from Definition 14.7 and (14.27) that $\Phi_i(t) = \Phi_i(u)/2$ and from formula (2.17) that $\gamma(t) = 2\gamma(u)$. Both order conditions are thus identical. □

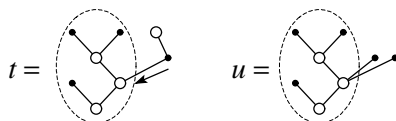


Fig. 14.2. Trees of Lemma 14.14

Condition (14.26) allows us to neglect the equations (14.23), while condition (14.27) plays a similar role to that of (1.9) for Runge-Kutta methods. It expresses the fact that the g_i of (14.13) approximate $y(x_0 + c_i h)$ up to $\mathcal{O}(h^3)$. As a consequence of Lemma 14.14, SN-trees which have at least one fat end-vertex can be left out (i.e., $t_4, t_6, t_9, t_{10}, t_{13}$ of Table 14.3).

With the help of (14.26) and (14.27) *explicit Nyström methods* (14.8) of order 5 with $s = 4$ can now easily be constructed: the order conditions for the trees t_1, t_2, t_3, t_5 and t_8 just indicate that the quadrature formula with nodes $c_1 = 0, c_2, c_3, c_4$ and weights b_1, b_2, b_3, b_4 is of order 5. Thus the nodes c_i have to satisfy the orthogonality relation

$$\int_0^1 x(x - c_2)(x - c_3)(x - c_4) dx = 0$$

and we see that two degrees of freedom are still left in the choice of the quadrature formula. The \bar{a}_{ij} are now uniquely determined and can be computed as follows: \bar{a}_{21} is given by (14.27) for $i = 2$. The order conditions for t_7 and t_{11} constitute two linear equations for the unknowns

$$\sum_{j=1}^2 \bar{a}_{3j} c_j \quad \text{and} \quad \sum_{j=1}^3 \bar{a}_{4j} c_j.$$

Together with (14.27, $i = 3$) one now obtains \bar{a}_{31} and \bar{a}_{32} . Finally, the order condition for t_{12} leads to $\sum_j \bar{a}_{4j} c_j^2$ and the remaining coefficients $\bar{a}_{41}, \bar{a}_{42}, \bar{a}_{43}$ can be computed from a Vandermonde-type linear system. The method of Table 14.2 is obtained in this way.

For still higher order methods it is helpful to use further simplifying assumptions; for example

$$\sum_{j=1}^s \bar{a}_{ij} c_j^q = \frac{c_i^{q+2}}{(q+2)(q+1)} \quad (14.28)$$

which, for $q = 0$, reduces to (14.27), and

$$\sum_{i=1}^s b_i c_i^q \bar{a}_{ij} = b_j \left(\frac{c_j^{q+2}}{(q+2)(q+1)} - \frac{c_j}{q+1} + \frac{1}{q+2} \right) \quad (14.29)$$

which can be considered a generalization of condition $D(\zeta)$ of Section II.7. For more details we refer to Hairer & Wanner (1976) and also to Albrecht (1955), Battin (1976), Beentjes & Gerritsen (1976), Hairer (1977, 1982), where Nyström methods of higher order are presented.

Embedded Nyström methods. For an efficient implementation we need a step size control mechanism. This can be performed in the same manner as for Runge-Kutta methods (see Section II.4). One can either apply Richardson extrapolation in order to estimate the local error, or construct embedded Nyström methods.

A series of embedded Nyström methods has been constructed by Fehlberg (1972). These methods use a $(p+1)$ -st order approximation to $y(x_0+h)$ for step size control. A $(p+1)$ -st order approximation to $y'(x_0+h)$ is not needed, since the lower order approximations are used for step continuation.

As for first order differential equations, local extrapolation — to use the higher order approximations for step continuation — turns out to be superior. Bettis (1973) was apparently the first to use this technique. His proposed method is of order 5(4). A method of order 7(6) has been constructed by Dormand & Prince (1978), methods of order 8(7), 9(8), 10(9) and 11(10) are given by Filippi & Gräf (1986) and further methods of order 8(6) and 12(10) are presented by Dormand, El-Mikkawy & Prince (1987).

In certain situations (see Section II.6) it is important that a Nyström method be equipped with a dense output formula. Such procedures are given by Dormand & Prince (1987) and, for general initial value problems $y'' = f(x, y, y')$, by Fine (1987).

An Extrapolation Method for $y'' = f(x, y)$

Les calculs originaux, comprenant environ 3.000 pages in-folio avec 358 grandes planches, et encore 3.800 pages de développements mathématiques correspondants, appartiennent maintenant à la collection de manuscrits de la Bibliothèque de l'Université, Christiania. (Störmer 1921)

If we rewrite the differential equation (14.7) as a first order system

$$\begin{pmatrix} y \\ y' \end{pmatrix}' = \begin{pmatrix} y' \\ f(x, y) \end{pmatrix}, \quad \begin{pmatrix} y \\ y' \end{pmatrix}(x_0) = \begin{pmatrix} y_0 \\ y'_0 \end{pmatrix} \quad (14.30)$$

we can apply the GBS-algorithm (9.13) directly to (14.30); this yields

$$y_1 = y_0 + hy'_0 \quad (14.31a)$$

$$y'_1 = y'_0 + hf(x_0, y_0)$$

$$y_{i+1} = y_{i-1} + 2hy'_i \quad (14.31b)$$

$$y'_{i+1} = y'_{i-1} + 2hf(x_i, y_i) \quad i = 1, 2, \dots, 2n$$

$$S_h(x) = (y_{2n-1} + 2y_{2n} + y_{2n+1})/4 \quad (14.31c)$$

$$S'_h(x) = (y'_{2n-1} + 2y'_{2n} + y'_{2n+1})/4.$$

Here, $S_h(x)$ and $S'_h(x)$ are the numerical approximations to $y(x)$ and $y'(x)$ at $x = x_0 + H$, where $H = 2nh$ and $x_i = x_0 + ih$. We now make the following important observation: for the computation of $y_0, y_2, y_4, \dots, y_{2n}$ (even indices) and

of $y'_1, y'_3, \dots, y'_{2n+1}$ (odd indices) only the function values $f(x_0, y_0), f(x_2, y_2), \dots, f(x_{2n}, y_{2n})$ have to be calculated. Furthermore, we know from (9.17) that y_{2n} and $(y'_{2n-1} + y'_{2n+1})/2$ each possess an asymptotic expansion in even powers of h . It is therefore obvious that (14.31c) should be replaced by (Gragg 1965)

$$\begin{aligned} S_h(x) &= y_{2n} \\ S'_h(x) &= (y'_{2n-1} + y'_{2n+1})/2. \end{aligned} \quad (14.31c')$$

Using this final step, the number of function evaluations is reduced by a factor of two. These numerical approximations can now be used for extrapolation. We take the harmonic sequence (9.8'), put

$$T_{i1} = S_h(x_0 + H), \quad T'_{i1} = S'_h(x_0 + H)$$

and compute the extrapolated expressions $T_{i,j}$ and $T'_{i,j}$ by the Aitken & Neville formula (9.10).

Remark. Eliminating the y'_j -values in (14.31b) we obtain the equivalent formula

$$y_{i+2} - 2y_i + y_{i-2} = (2h)^2 f(x_i, y_i), \quad (14.32)$$

which is often called *Störmer's rule*. For the implementation the formulation (14.31b) is to be preferred, since it is more stable with respect to round-off errors (see Section III.10).

Dense output. As for the derivation of Section II.9 for the GBS algorithm we shall do Hermite interpolation based on derivatives of the solution at $x_0, x_0 + H$ and $x_0 + H/2$. At the endpoints of the considered interval we have $y_0, y'_0, y''_0 = f(x_0, y_0)$ and y_1, y'_1, y''_1 at our disposal. The derivatives at the midpoint can be obtained by extrapolation of suitable differences of function values. However, one has to take care of the fact that y_i and $f(x_i, y_i)$ are available only for even indices, whereas y'_i is available for odd indices only. For the same reason as for the GBS method, the step number sequence has to satisfy (9.34). For notational convenience, the following description is restricted to the sequence (9.35).

We suppose that T_{kk} and T'_{kk} are accepted approximations to the solution. Then the construction of a dense output formula can be summarized as follows:

Step 1. For each $j \in \{1, \dots, k\}$ compute the approximations to the derivatives of $y(x)$ at $x_0 + H/2$ by (δ is the central difference operator):

$$\begin{aligned} d_j^{(0)} &= \frac{1}{2} (y_{n_j/2-1} + y_{n_j/2+1}), & d_j^{(1)} &= y'_{n_j/2}, \\ d_j^{(\kappa)} &= \frac{1}{2} \cdot \frac{1}{(2h_j)^{\kappa-2}} \left(\delta^{\kappa-2} f_{n_j/2-1}^{(j)} + \delta^{\kappa-2} f_{n_j/2+1}^{(j)} \right), & \kappa &= 2, 4, \dots, 2j, \\ d_j^{(\kappa)} &= \frac{\delta^{\kappa-2} f_{n_j/2}^{(j)}}{(2h_j)^{\kappa-2}}, & \kappa &= 3, 5, \dots, 2j+1. \end{aligned} \quad (14.33)$$

Step 2. Extrapolate $d_j^{(0)}, d_j^{(1)}$ ($k-1$) times and $d_j^{(2\ell)}, d_j^{(2\ell+1)}$ ($k-\ell$) times to obtain improved approximations $d^{(\kappa)}$ to $y^{(\kappa)}(x_0 + H/2)$.

Step 3. For given μ ($-1 \leq \mu \leq 2k+1$) define the polynomial $P_\mu(\theta)$ of degree $\mu+6$ by

$$\begin{aligned} P_\mu(0) &= y_0, & P'_\mu(0) &= y'_0, & P''_\mu(0) &= f(x_0, y_0) \\ P_\mu(1) &= T_{kk}, & P'_\mu(1) &= T'_{kk}, & P''_\mu(1) &= f(x_0 + H, T_{kk}) \\ P_\mu^{(\kappa)}(1/2) &= H^\kappa d^{(\kappa)} & & \text{for } \kappa = 0, 1, \dots, \mu. \end{aligned} \quad (14.34)$$

Since T_{kk}, T'_{kk} are the initial values for the next step, the dense output obtained by the above algorithm is a global \mathcal{C}^2 approximation to the solution. It satisfies

$$y(x_0 + \theta H) - P_\mu(\theta) = \mathcal{O}(H^{2k}) \quad \text{if } \mu \geq 2k-7 \quad (14.35)$$

(compare Theorem 9.5). In the code ODEX2 of the Appendix the value $\mu = 2k-5$ is suggested as standard choice.

Problems for Numerical Comparisons

PLEI — the celestial mechanics problem (10.3) which is the only problem of Section II.10 already in the special form (14.7).

ARES — the AREnstorf orbit in Second order form (14.7). This is the restricted three body problem (0.1) with initial values (0.2) integrated over one period $0 \leq x \leq x_{\text{end}}$ (see Fig. 0.1) in a *fixed* coordinate system. Then the equations of motion become

$$\begin{aligned} y_1'' &= \mu' \frac{a_1(x) - y_1}{D_1} + \mu \frac{b_1(x) - y_1}{D_2} \\ y_2'' &= \mu' \frac{a_2(x) - y_2}{D_1} + \mu \frac{b_2(x) - y_2}{D_2} \end{aligned} \quad (14.36)$$

where

$$D_1 = ((y_1 - a_1(x))^2 + (y_2 - a_2(x))^2)^{3/2}, \quad D_2 = ((y_1 - b_1(x))^2 + (y_2 - b_2(x))^2)^{3/2}$$

and the movement of sun and moon are described by

$$a_1(x) = -\mu \cos x \quad a_2(x) = -\mu \sin x \quad b_1(x) = \mu' \cos x \quad b_2(x) = \mu' \sin x.$$

The initial values

$$\begin{aligned} y_1(0) &= 0.994, & y_1'(0) &= 0, & y_2(0) &= 0, \\ y_2'(0) &= -2.00158510637908252240537862224 + 0.994, \\ x_{\text{end}} &= 17.0652165601579625588917206249, \end{aligned}$$

are those of (0.2) enlarged by the speed of the rotation. The exact solution values are the initial values transformed by the rotation of the coordinate system.

CPEN — the nonlinear Coupled PENDulum (see Fig. 14.3).

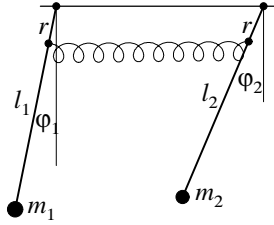


Fig. 14.3. Coupled pendulum

The kinetic as well as potential energies

$$\begin{aligned} T &= \frac{m_1 l_1^2 \dot{\varphi}_1^2}{2} + \frac{m_2 l_2^2 \dot{\varphi}_2^2}{2} \\ V &= -m_1 l_1 \cos \varphi_1 - m_2 l_2 \cos \varphi_2 + \frac{c_0 r^2 (\sin \varphi_1 - \sin \varphi_2)^2}{2} \end{aligned}$$

lead by Lagrange theory (equations (I.6.21)) to

$$\begin{aligned} \ddot{\varphi}_1 &= -\frac{\sin \varphi_1}{l_1} - \frac{c_0 r^2}{m_1 l_1^2} (\sin \varphi_1 - \sin \varphi_2) \cos \varphi_1 + f(t) \\ \ddot{\varphi}_2 &= -\frac{\sin \varphi_2}{l_2} - \frac{c_0 r^2}{m_2 l_1^2} (\sin \varphi_2 - \sin \varphi_1) \cos \varphi_2. \end{aligned} \tag{14.37}$$

We choose the parameters

$$l_1 = l_2 = 1, \quad m_1 = 1, \quad m_2 = 0.99, \quad r = 0.1, \quad c_0 = 0.01, \quad t_{\text{end}} = 496$$

and all initial values and speeds for $t = 0$ equal to zero. The first pendulum is then pushed into movement by a (somewhat idealized) hammer as

$$f(t) = \begin{cases} \sqrt{1 - (1 - t)^2} & \text{if } |t - 1| \leq 1; \\ 0 & \text{otherwise.} \end{cases}$$

The resulting solutions are displayed in Fig. 14.4. The nonlinearities in this problem produce quite different sausages (cf. “Mon Oncle” de Jacques Tati 1958) from those people are accustomed to from linear problems (cf. Sommerfeld 1942, §20).

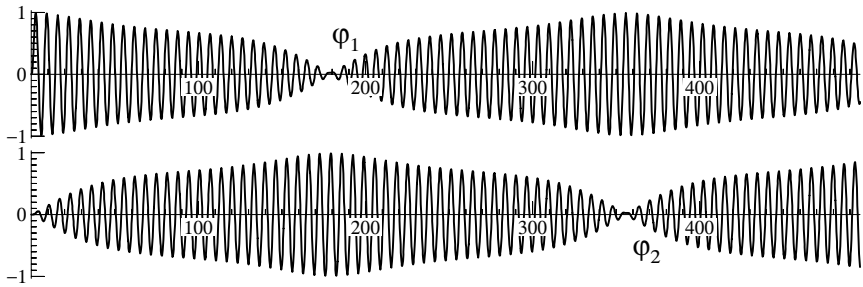


Fig. 14.4. Movement of the coupled pendulum (14.37)

WPLT — the Weak PLaTe, i.e., the PLATE problem of Section IV.10 (see Volume II) with weakened stiffness. We use precisely the same equations as (IV.10.6) and reduce the stiffness parameter σ from $\sigma = 100$ to $\sigma = 1/16$. We also remove the friction ($\omega = 0$ instead of $\omega = 1000$) so that the problem becomes purely of second order. It is linear, nonautonomous, and of dimension 40.

Performance of the Codes

Several codes were applied to each of the above four problems with 89 different tolerances between $Tol = 10^{-3}$ and $Tol = 10^{-14}$ (exactly as in Section II.10). The number of function evaluations (Fig. 14.5) and the computer time (Fig. 14.6) on a Sun Workstation (SunBlade 100) are plotted as a function of the global error at the endpoint of the integration interval. The codes used are the following:

RKN6 — symbol \star — is the low order option of the Runge-Kutta-Nyström code presented in Brankin, Gladwell, Dormand, Prince & Seward (1989). It is based on a fixed-order embedded Nyström method of order 6(4), whose coefficients are given in Dormand & Prince (1987). This code is provided with a dense output.

RKN12 — symbol \boxtimes — is the high order option of the Runge-Kutta-Nyström code presented in Brankin & al. (1989). It is based on the method of order 12(10), whose coefficients are given in Dormand, El-Mikkawy & Prince (1987). This code is not equipped with a dense output.

ODEX2 — symbol \bigcirc — is the extrapolation code based on formula (14.31a,b,c') and uses the harmonic step number sequence (see Appendix). It is implemented in the same way as ODEX (the extrapolation code for first order differential equations). In particular, the order and step size strategy is that of Section II.9. A dense output is available. Similar results are obtained by the code DIFEX2 of Deuffhard & Bauer (see Deuffhard 1985).

In order to demonstrate the superiority of the special methods for $y'' = f(x, y)$, we have included the results obtained by DOP853 (symbol \star) and ODEX (symbol

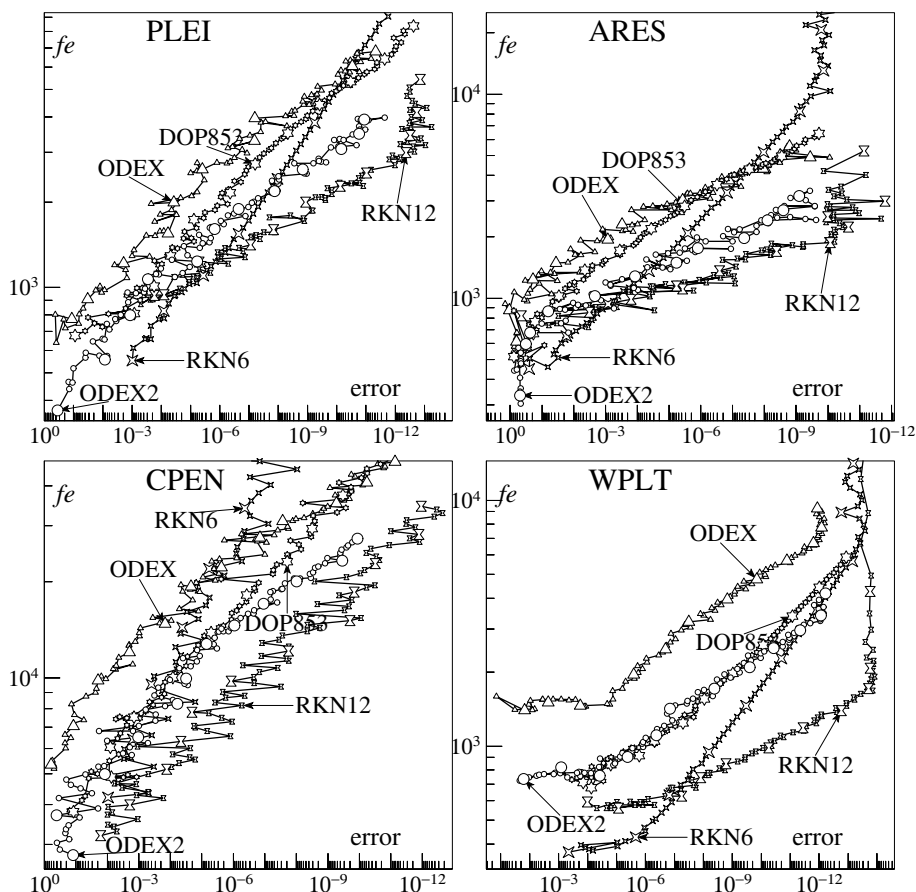


Fig. 14.5. Precision versus function evaluations

\triangle) which were already described in Section II.10. For their application we had to rewrite the four problems as a first order system by introducing the first derivatives as new variables. The code ODEX2 is nearly twice as efficient as ODEX which is in agreement with the theoretical considerations. Similarly the Runge-Kutta-Nyström codes RKN6 and RKN12 are a real improvement over DOP853.

A comparison of Fig. 14.5 and 14.6 shows a significant difference. The extrapolation codes ODEX and ODEX2 are relatively better on the “time”-pictures than for the function evaluation counts. With the exception of problem WPLT the performance of the code ODEX2 then becomes comparable to that of RKN12. As can be observed especially at the WPLT problem, the code RKN12 overshoots, for stringent tolerances, significantly the desired precision. It becomes less efficient if Tol is chosen too close to $Uround$.

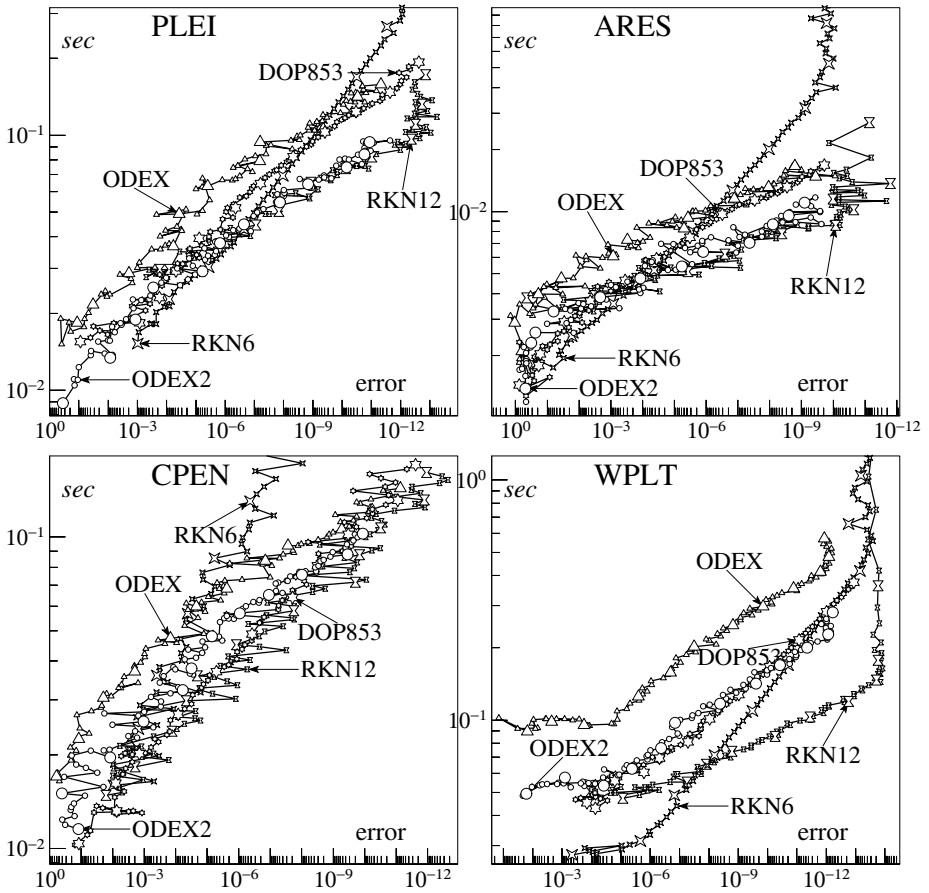


Fig. 14.6. Precision versus computing time

Exercises

1. Verify that the methods of Table 14.2 are of order 4 and 5, respectively.
2. The error coefficients of a p th order Nyström method are defined by

$$\begin{aligned} e(t) &= 1 - (\varrho(t) + 1)\gamma(t) \sum_i \bar{b}_i \Phi_i(t) & \text{for } \varrho(t) = p, \\ e'(t) &= 1 - \gamma(t) \sum_i b_i \Phi_i(t) & \text{for } \varrho(t) = p + 1. \end{aligned} \quad (14.38)$$

- a) The assumption (14.26) implies that

$$e(t) = -\varrho(t)e'(u) \quad \text{for } \varrho(t) = p,$$

where u is the N-tree obtained from t by adding a branch with a meagre vertex to the root of t .

- b) Compute the error coefficients of Nyström's method (Table 14.1) and compare them to those of the classical Runge-Kutta method.
3. Show that the order conditions for Runge-Kutta methods (Theorem 2.13) are a subset of the conditions (14.21). They correspond to the N-trees, all of whose vertices are fat.
4. Sometimes the definition of order of Nyström methods (14.8) is relaxed to

$$\begin{aligned} y(x_0 + h) - y_1 &= \mathcal{O}(h^{p+1}) \\ y'(x_0 + h) - y'_1 &= \mathcal{O}(h^p) \end{aligned} \quad (14.39)$$

(see Nyström 1925). Show that the conditions (14.39) are not sufficient to obtain global convergence of order p .

Hint. Investigate the asymptotic expansion of the global error with the help of Theorem 8.1 and formula (8.8).

5. The numerical solutions T_{kk} and T'_{kk} of the extrapolation method of this section are equivalent to a Nyström method of order $p = 2k$ with $s = p^2/8 + p/4 + 1$ stages.
6. A *collocation method* for $y'' = f(x, y, y')$ (or $y'' = f(x, y)$) can be defined as follows: let $u(x)$ be a polynomial of degree $s + 1$ defined by

$$\begin{aligned} u(x_0) &= y_0, & u'(x_0) &= y'_0 \\ u''(x_0 + c_i h) &= f(x_0 + c_i h, u(x_0 + c_i h), u'(x_0 + c_i h)), & i &= 1, \dots, s, \end{aligned} \quad (14.40)$$

then the numerical solution is given by $y_1 = u(x_0 + h)$, $y'_1 = u'(x_0 + h)$.

- a) Prove that this collocation method is equivalent to the Nyström method (14.4) where

$$\begin{aligned} a_{ij} &= \int_0^{c_i} \ell_j(t) dt, & \bar{a}_{ij} &= \int_0^{c_i} (c_i - t) \ell_j(t) dt, \\ b_i &= \int_0^1 \ell_i(t) dt, & \bar{b}_i &= \int_0^1 (1 - t) \ell_i(t) dt, \end{aligned} \quad (14.41)$$

and $\ell_j(t)$ are the Lagrange polynomials of (7.17).

- b) The a_{ij} satisfy $C(s)$ (see Theorem 7.8) and the \bar{a}_{ij} satisfy (14.28) for $q = 0, 1, \dots, s - 1$. These equations uniquely define a_{ij} and \bar{a}_{ij} .
- c) In general, a_{ij} and \bar{a}_{ij} do not satisfy (14.5).
- d) If $M(t) = \prod_{i=1}^s (t - c_i)$ is orthogonal to all polynomials of degree $r - 1$,

$$\int_0^1 M(t) t^{q-1} dt = 0, \quad q = 1, \dots, r,$$

then the collocation method (14.40) has order $p = s + r$.

- e) The polynomial $u(x)$ yields an approximation to the solution $y(x)$ on the whole interval $[x_0, x_0 + h]$. The following estimates hold:

$$y(x) - u(x) = \mathcal{O}(h^{s+2}), \quad y'(x) - u'(x) = \mathcal{O}(h^{s+1}).$$

II.15 P-Series for Partitioned Differential Equations

Divide ut regnes

(N. Machiavelli 1469-1527)

In the previous section we considered direct methods for second order differential equations $y'' = f(y, y')$. The idea was to write the equation as a partitioned differential system

$$\begin{pmatrix} y \\ y' \end{pmatrix}' = \begin{pmatrix} y' \\ f(y, y') \end{pmatrix} \quad (15.1)$$

and to discretize the two components, y and y' , by different formulas. There are many other situations where the problem possesses a natural partitioning. Typical examples are the Hamiltonian equations (I.6.26, I.14.26) and singular perturbation problems (see Chapter VI of Volume II). It may also be of interest to separate linear and nonlinear parts or the “non-stiff” and “stiff” components of a differential equation.

We suppose that the differential system is partitioned as

$$\begin{pmatrix} y_a \\ y_b \end{pmatrix}' = \begin{pmatrix} f_a(y_a, y_b) \\ f_b(y_a, y_b) \end{pmatrix} \quad (15.2)$$

where the solution vector is separated into two components y_a, y_b , each of which may itself be a vector. An extension to more components is straight-forward.

For the numerical solution of (15.2) we consider the *partitioned method*

$$\begin{aligned} k_i &= f_a\left(y_{a0} + h \sum_{j=1}^s a_{ij} k_j, y_{b0} + h \sum_{j=1}^s \hat{a}_{ij} \ell_j\right) \\ \ell_i &= f_b\left(y_{a0} + h \sum_{j=1}^s a_{ij} k_j, y_{b0} + h \sum_{j=1}^s \hat{a}_{ij} \ell_j\right) \\ y_{a1} &= y_{a0} + h \sum_{i=1}^s b_i k_i, \quad y_{b1} = y_{b0} + h \sum_{i=1}^s \hat{b}_i \ell_i \end{aligned} \quad (15.3)$$

where the coefficients a_{ij}, b_i and \hat{a}_{ij}, \hat{b}_i represent two different Runge-Kutta schemes. The first methods of this type are due to Hofer (1976) and Griepentrog (1978) who apply an explicit method to the nonstiff part and an implicit method to the stiff part of a differential equation. Later Rentrop (1985) modified this idea by combining explicit Runge-Kutta methods with Rosenbrock-type methods (Sec-

tion IV.7). Recent interest for partitioned methods came up when solving Hamiltonian systems (see Section II.16 below).

The subject of this section is the derivation of the order conditions for method (15.3). For order p it is necessary that each of the two Runge-Kutta schemes under consideration be of order p . This can be seen by applying the method to $y'_a = f_a(y_a)$, $y'_b = f_b(y_b)$. But this is not sufficient, the coefficients have to satisfy certain *coupling conditions*. In order to understand this, we first look at the derivatives of the exact solution of (15.2). Then we generalize the theory of B-series (see Section II.12) to the new situation (Hairer 1981) and derive the order conditions in the same way as in II.12 for Runge-Kutta methods.

Derivatives of the Exact Solution, P-Trees

In order to avoid sums and unnecessary indices we assume that y_a and y_b in (15.2) are scalar quantities. All subsequent formulas remain valid for vectors if the derivatives are interpreted as multi-linear mappings. Differentiating (15.2) and inserting (15.2) again for the derivatives we obtain for the first component y_a

$$y_a^{(1)} = f_a \quad (15.4;1)$$

$$y_a^{(2)} = \frac{\partial f_a}{\partial y_a} f_a + \frac{\partial f_a}{\partial y_b} f_b \quad (15.4;2)$$

$$y_a^{(3)} = \frac{\partial^2 f_a}{\partial y_a^2} (f_a, f_a) + \frac{\partial^2 f_a}{\partial y_b \partial y_a} (f_b, f_a) + \frac{\partial f_a}{\partial y_a} \frac{\partial f_a}{\partial y_a} f_a + \frac{\partial f_a}{\partial y_a} \frac{\partial f_a}{\partial y_b} f_b \\ + \frac{\partial^2 f_a}{\partial y_a \partial y_b} (f_a, f_b) + \frac{\partial^2 f_a}{\partial y_b^2} (f_b, f_b) + \frac{\partial f_a}{\partial y_b} \frac{\partial f_b}{\partial y_a} f_a + \frac{\partial f_a}{\partial y_b} \frac{\partial f_b}{\partial y_b} f_b. \quad (15.4;3)$$

Similar formulas hold for the derivatives of y_b .

For a graphical representation of these formulas we need two different kinds of vertices. As in Section II.14 we use “meagre” and “fat” vertices, which will correspond to f_a and f_b , respectively. Formulas (15.4) can then be represented as shown in Fig. 15.1.

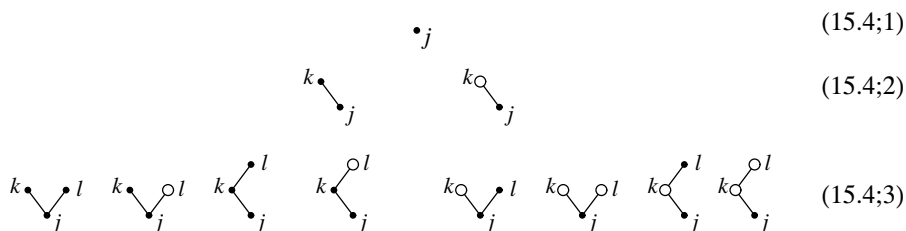


Fig. 15.1. The derivatives of the exact solution y_a

Definition 15.1. A *labelled P-tree* of order q is a labelled tree (see Definition 2.2)

$$t : A_q \setminus \{j\} \rightarrow A_q$$

together with a mapping

$$t' : A_q \rightarrow \{\text{“meagre”}, \text{“fat”}\}.$$

We denote by LP_q^a the set of those labelled P-trees of order q , whose root is meagre (i.e., $t'(j) = \text{“meagre”}$). Similarly, LP_q^b is the set of q th order labelled P-trees with a “fat” root.

Due to the symmetry of the second derivative the 2nd and 5th expressions in (15.4;3) are equal. We therefore define:

Definition 15.2. Two labelled P-trees (t, t') and (u, u') are *equivalent*, if they have the same order, say q , and if there exists a bijection $\sigma : A_q \rightarrow A_q$ such that $\sigma(j) = j$ and the following diagram commutes:

$$\begin{array}{ccccc} A_q \setminus \{j\} & \xrightarrow{t} & A_q & \xrightarrow{t'} & \{\text{“meagre”}, \text{“fat”}\} \\ \sigma \downarrow & & \sigma \downarrow & \nearrow & \\ A_q \setminus \{j\} & \xrightarrow{u} & A_q & \nearrow u' & \end{array}$$

Definition 15.3. An equivalence class of q th order labelled P-trees is called a *P-tree* of order q . The set of all P-trees of order q with a meagre root is denoted by TP_q^a , that with a fat root by TP_q^b . For a P-tree t we denote by $\varrho(t)$ the *order* of t , and by $\alpha(t)$ the number of elements in the equivalence class t .

Examples of P-trees together with the numbers $\varrho(t)$ and $\alpha(t)$ are given in Table 15.1 below. We first discuss a recursive representation of P-trees (extension of Definition 2.12), which is fundamental for the following theory.

Definition 15.4. Let t_1, \dots, t_m be P-trees. We then denote by

$$t = {}_a[t_1, \dots, t_m] \quad (15.5)$$

the unique P-tree t such that the root is “meagre” and the P-trees t_1, \dots, t_m remain if the root and the adjacent branches are chopped off. Similarly, we denote by ${}_b[t_1, \dots, t_m]$ the P-tree whose new root is “fat” (see Fig. 15.2). We further denote by τ_a and τ_b the meagre and fat P-trees of order one.

Our next aim is to make precise the connection between P-trees and the expressions of the formulas (15.4). For this we use the notation

$$w(t) = \begin{cases} a & \text{if the root of } t \text{ is meagre,} \\ b & \text{if the root of } t \text{ is fat.} \end{cases} \quad (15.6)$$

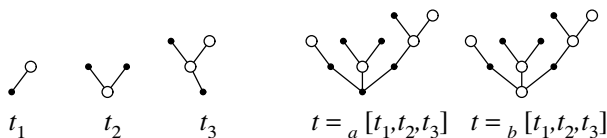


Fig. 15.2. Recursive definition of P-trees

Definition 15.5. The *elementary differentials*, corresponding to (15.2), are defined recursively by $(y = (y_a, y_b))$

$$F(\tau_a)(y) = f_a(y), \quad F(\tau_b)(y) = f_b(y)$$

and

$$F(t)(y) = \frac{\partial^m f_{w(t)}(y)}{\partial y_{w(t_1)} \dots \partial y_{w(t_m)}} \cdot (F(t_1)(y), \dots, F(t_m)(y))$$

for $t = {}_a[t_1, \dots, t_m]$ or $t = {}_b[t_1, \dots, t_m]$.

Elementary differentials for P-trees up to order 3 are given explicitly in Table 15.1.

We now return to the starting-point of this section and continue the differentiation of formulas (15.4). Using the notation of labelled P-trees, one sees that a differentiation of $F(t)(y_a, y_b)$ can be interpreted as an addition of a new branch with a meagre or fat vertex and a new summation letter to each vertex of the labelled P-tree t . In the same way as we proved Theorem 2.6 for non-partitioned differential equations, we arrive at

Theorem 15.6. *The derivatives of the exact solution of (15.2) satisfy*

$$\begin{aligned} y_a^{(q)} &= \sum_{t \in LTP_q^a} F(t)(y_a, y_b) = \sum_{t \in TTP_q^a} \alpha(t) F(t)(y_a, y_b) \\ y_b^{(q)} &= \sum_{t \in LTP_q^b} F(t)(y_a, y_b) = \sum_{t \in TTP_q^b} \alpha(t) F(t)(y_a, y_b). \end{aligned} \quad (15.4;q)$$

□

Table 15.1. P-trees and their elementary differentials

P-tree	repr. (15.5)	$\varrho(t)$	$\alpha(t)$	elem. differential	$\Phi_j(t)$
	τ_a	1	1	f_a	1
	$a[\tau_a]$	2	1	$\frac{\partial f_a}{\partial y_a} f_a$	$\sum_k a_{jk}$
	$a[\tau_b]$	2	1	$\frac{\partial f_a}{\partial y_b} f_b$	$\sum_k \hat{a}_{jk}$
	$a[\tau_a, \tau_a]$	3	1	$\frac{\partial^2 f_a}{\partial y_a^2} (f_a, f_a)$	$\sum_{k,l} a_{jk} a_{jl}$
	$a[\tau_a, \tau_b]$	3	2	$\frac{\partial^2 f_a}{\partial y_a \partial y_b} (f_a, f_b)$	$\sum_{k,l} a_{jk} \hat{a}_{jl}$
	$a[\tau_b, \tau_b]$	3	1	$\frac{\partial^2 f_a}{\partial y_b^2} (f_b, f_b)$	$\sum_{k,l} \hat{a}_{jk} \hat{a}_{jl}$
	$a[a[\tau_a]]$	3	1	$\frac{\partial f_a}{\partial y_a} \frac{\partial f_a}{\partial y_a} f_a$	$\sum_{k,l} a_{jk} a_{kl}$
	$a[a[\tau_b]]$	3	1	$\frac{\partial f_a}{\partial y_a} \frac{\partial f_a}{\partial y_b} f_b$	$\sum_{k,l} a_{jk} \hat{a}_{kl}$
	$a[b[\tau_a]]$	3	1	$\frac{\partial f_a}{\partial y_b} \frac{\partial f_b}{\partial y_a} f_a$	$\sum_{k,l} \hat{a}_{jk} a_{kl}$
	$a[b[\tau_b]]$	3	1	$\frac{\partial f_a}{\partial y_b} \frac{\partial f_b}{\partial y_b} f_b$	$\sum_{k,l} \hat{a}_{jk} \hat{a}_{kl}$
...
	τ_b	1	1	f_b	1
	$b[\tau_a]$	2	1	$\frac{\partial f_b}{\partial y_a} f_a$	$\sum_k a_{jk}$
	$b[\tau_b]$	2	1	$\frac{\partial f_b}{\partial y_b} f_b$	$\sum_k \hat{a}_{jk}$
...

P-Series

In Section II.12 we saw the importance of the key-lemma Corollary 12.7 for the derivation of the order conditions for Runge-Kutta methods. Therefore we extend this result also to partitioned ordinary differential equations.

It is convenient to introduce two new P-trees of order 0, namely \emptyset_a and \emptyset_b . The corresponding elementary differentials are $F(\emptyset_a)(y) = y_a$ and $F(\emptyset_b)(y) = y_b$. We further set

$$\begin{aligned}
 TP^a &= \{\emptyset_a\} \cup TP_1^a \cup TP_2^a \cup \dots & LTP^a &= \{\emptyset_a\} \cup LTP_1^a \cup LTP_2^a \cup \dots \\
 TP^b &= \{\emptyset_b\} \cup TP_1^b \cup TP_2^b \cup \dots & LTP^b &= \{\emptyset_b\} \cup LTP_1^b \cup LTP_2^b \cup \dots
 \end{aligned}
 \tag{15.7}$$

Definition 15.7. Let $\mathbf{c}(\emptyset_a)$, $\mathbf{c}(\emptyset_b)$, $\mathbf{c}(\tau_a)$, $\mathbf{c}(\tau_b)$, ... be real coefficients defined for all P-trees, i.e., $\mathbf{c} : T P^a \cup T P^b \rightarrow \mathbb{R}$. The series

$$P(\mathbf{c}, y) = (P_a(\mathbf{c}, y), P_b(\mathbf{c}, y))^T$$

where

$$P_a(\mathbf{c}, y) = \sum_{t \in LTP^a} \frac{h^{\varrho(t)}}{\varrho(t)!} \mathbf{c}(t) F(t)(y), \quad P_b(\mathbf{c}, y) = \sum_{t \in LTP^b} \frac{h^{\varrho(t)}}{\varrho(t)!} \mathbf{c}(t) F(t)(y)$$

is then called a *P-series*.

Theorem 15.6 simply states that the exact solution of (15.2) is a P-series

$$(y_a(x_0 + h), y_b(x_0 + h))^T = P(\mathbf{y}, (y_a(x_0), y_b(x_0))) \quad (15.8)$$

with $\mathbf{y}(t) = 1$ for all P-trees t .

Theorem 15.8. Let $\mathbf{c} : T P^a \cup T P^b \rightarrow \mathbb{R}$ be a sequence of coefficients such that $\mathbf{c}(\emptyset_a) = \mathbf{c}(\emptyset_b) = 1$. Then

$$h \begin{pmatrix} f_a(P(\mathbf{c}, (y_a, y_b))) \\ f_b(P(\mathbf{c}, (y_a, y_b))) \end{pmatrix} = P(\mathbf{c}', (y_a, y_b)) \quad (15.9)$$

with

$$\begin{aligned} \mathbf{c}'(\emptyset_a) = \mathbf{c}'(\emptyset_b) &= 0, & \mathbf{c}'(\tau_a) = \mathbf{c}'(\tau_b) &= 1 \\ \mathbf{c}'(t) = \varrho(t) \mathbf{c}(t_1) \dots \mathbf{c}(t_m) & \quad \text{if } t = {}_a[t_1, \dots, t_m] \text{ or } t = {}_b[t_1, \dots, t_m]. \end{aligned} \quad (15.10)$$

The *proof* is related to that of Theorem 12.6. It is given with more details in Hairer (1981). \square

Order Conditions for Partitioned Runge-Kutta Methods

With the help of Theorem 15.8 the order conditions for method (15.3) can readily be obtained. For this we denote the arguments in (15.3) by

$$g_i = y_{a0} + h \sum_{j=1}^s a_{ij} k_j, \quad \widehat{g}_i = y_{b0} + h \sum_{j=1}^s \widehat{a}_{ij} \ell_j, \quad (15.11)$$

and we assume that $G_i = (g_i, \widehat{g}_i)^T$ and $K_i = h(k_i, \ell_i)^T$ are P-series with coefficients $\mathbf{G}_i(t)$ and $\mathbf{K}_i(t)$, respectively. The formulas (15.11) then yield $\mathbf{G}_i(\emptyset_a) = 1$, $\mathbf{G}_i(\emptyset_b) = 1$ and

$$\mathbf{G}_i(t) = \begin{cases} \sum_{j=1}^s a_{ij} \mathbf{K}_j(t) & \text{if the root of } t \text{ is meagre,} \\ \sum_{j=1}^s \widehat{a}_{ij} \mathbf{K}_j(t) & \text{if the root of } t \text{ is fat.} \end{cases} \quad (15.12)$$

Application of Theorem 15.8 to the relations $k_j = f_a(G_j)$, $\ell_j = f_b(G_j)$ shows that $\mathbf{K}_j(t) = \mathbf{G}'_j(t)$ which, together with (15.10) and (15.12), recursively defines the values $\mathbf{K}_j(t)$.

It is usual to write $\mathbf{K}_j(t) = \gamma(t)\Phi_j(t)$ where $\gamma(t)$ is the integer given in Definition 2.10 (see also (2.17)). The coefficient $\Phi_j(t)$ is then obtained in the same way as the corresponding value of standard Runge-Kutta methods (see Definition 2.9) with the exception that a factor a_{ik} has to be replaced by \hat{a}_{ik} , if the vertex with label “ k ” is fat. A comparison of the P-series for the numerical solution $(y_{1a}, y_{1b})^T$ with that for the exact solution (15.8) yields the desired order conditions.

Theorem 15.9. *A partitioned Runge-Kutta method (15.3) is of order p iff*

$$\sum_{j=1}^s b_j \Phi_j(t) = \frac{1}{\gamma(t)} \quad \text{and} \quad \sum_{j=1}^s \hat{b}_j \Phi_j(t) = \frac{1}{\gamma(t)} \quad (15.13)$$

for all P-trees of order $\leq p$. □

Example. A partitioned method (15.3) is of order 2, if and only if each of the two Runge-Kutta schemes has order 2 and if the coupling conditions

$$\sum_{i,j} b_i \hat{a}_{ij} = \frac{1}{2}, \quad \sum_{i,j} \hat{b}_i a_{ij} = \frac{1}{2},$$

which correspond to trees ${}_a[\tau_b]$ and ${}_b[\tau_a]$ of Table 15.1 respectively, are satisfied. This happens if

$$c_i = \hat{c}_i \quad \text{for all } i.$$

This last assumption simplifies the order conditions considerably (the “thickness” of terminating vertices then has no influence). The resulting conditions for order up to 4 have been tabulated by Griepentrog (1978).

Further Applications of P-Series

Runge-Kutta methods violating (1.9). For the non-autonomous differential equation $y' = f(x, y)$ we consider, as in Exercise 6 of Section II.1, the Runge-Kutta method

$$k_i = f\left(x_0 + \hat{c}_i h, y_0 + h \sum_{j=1}^s a_{ij} k_j\right), \quad y_1 = y_0 + h \sum_{i=1}^s b_i k_i, \quad (15.14)$$

where \widehat{c}_i is not necessarily equal to $c_i = \sum_j a_{ij}$. Therefore, the x and y components in

$$\begin{aligned} y' &= f(x, y) \\ x' &= 1. \end{aligned} \quad (15.15)$$

are integrated differently. This system is of the form (15.2), if we put $y_a = y$, $y_b = x$, $f_a(y_a, y_b) = f(x, y)$ and $f_b(y_a, y_b) = 1$. Since f_b is constant, all elementary differentials that involve derivatives of f_b vanish identically. Thus, P-trees where at least one fat vertex is not an end-vertex need not be considered. It remains to treat the set

$$T_x = \{t \in TP_a; \text{ all fat vertices are end-vertices}\}. \quad (15.16)$$

Each tree of T_x gives rise to an order condition which is exactly that of Theorem 15.9. It is obtained in the usual way (Section II.2) with the exception that c_k has to be replaced by \widehat{c}_k , if the corresponding vertex is a fat one.

Fehlberg methods. The methods of Fehlberg, introduced in Section II.13, are equivalent to (15.14). However, it is known that the exact solution of the differential equation $y' = f(x, y)$ satisfies $y(x_0) = 0$, $y'(x_0) = 0, \dots, y^{(m)}(x_0) = 0$ at the initial value $x = x_0$. As explained in II.13, this implies that the expressions $f, \partial f / \partial x, \dots, \partial^{m-1} f / \partial x^{m-1}$ vanish at (x_0, y_0) and consequently also many of the elementary differentials disappear. The elements of T_x which remain to be considered are given in Fig. 15.3.

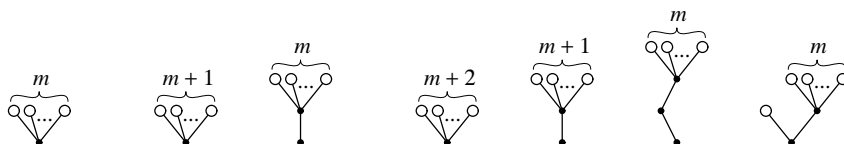


Fig. 15.3. P-trees for the methods of Fehlberg

Nyström methods. As a last application of Theorem 15.8 we present a new derivation of the order conditions for Nyström methods (Section II.14). The second order differential equation $y'' = f(y, y')$ can be written in partitioned form as

$$\begin{pmatrix} y \\ y' \end{pmatrix}' = \begin{pmatrix} y' \\ f(y, y') \end{pmatrix}. \quad (15.17)$$

In the notation of (15.2) we have $y_a = y$, $y_b = y'$, $f_a(y_a, y_b) = y_b$, $f_b(y_a, y_b) = f(y_a, y_b)$. The special structure of f_a implies that only P-trees which satisfy the condition (see Definition 14.2)

$$\text{“meagre vertices have at most one son and this son has to be fat”} \quad (15.18)$$

have to be considered. The essential P-trees are thus

$$TN_q^a = \{t \in TP_q^a ; t \text{ satisfies (15.18)}\}$$

$$TN_q^b = \{t \in TP_q^b ; t \text{ satisfies (15.18)}\}.$$

It follows that each element of TN_{q+1}^a can be written as $t = {}_a[u]$ with $u \in TN_q^b$. This implies a one-to-one correspondence between TN_{q+1}^a and TN_q^b , leaving the elementary differentials invariant:

$$F({}_a[u])(y_a, y_b) = \frac{\partial y_b}{\partial y_b} \cdot F(u)(y_a, y_b) = F(u)(y_a, y_b).$$

From this property it follows that

$$hP_b(\mathbf{c}, (y_a, y_b)) = P_a(\mathbf{c}', (y_a, y_b)) \quad (15.19)$$

where $\mathbf{c}'(\emptyset_a) = 0$, $\mathbf{c}'(\tau_a) = \mathbf{c}(\emptyset_b)$ and

$$\mathbf{c}'(t) = \varrho(t)\mathbf{c}(u) \quad \text{if } t = {}_a[u]. \quad (15.20)$$

This notation is in agreement with (15.10).

The order conditions of method (14.13) can now be derived as follows: assume g_i, g'_i to be P-series

$$g_i = P_a(\mathbf{c}_i, (y_0, y'_0)), \quad g'_i = P_b(\mathbf{c}_i, (y_0, y'_0)).$$

Theorem 15.8 then implies that

$$hf(g_i, g'_i) = P_b(\mathbf{c}'_i, (y_0, y'_0)). \quad (15.21)$$

Multiplying this relation by h it follows from (15.19) that

$$h^2 f(g_i, g'_i) = P_a(\mathbf{c}''_i, (y_0, y'_0)). \quad (15.22)$$

Here $\mathbf{c}''_i = (\mathbf{c}'_i)'$, i.e.,

$$\begin{aligned} \mathbf{c}''_i(t) &= 0 \quad \text{for } t = \emptyset_a \text{ and } t = \tau_a, & \mathbf{c}''_i({}_a[\tau_b]) &= 1, \\ \mathbf{c}''_i(t) &= \varrho(t)(\varrho(t) - 1)\mathbf{c}_i(t_1) \dots \mathbf{c}_i(t_m) & \text{if } t = {}_a[b[t_1, \dots, t_m]]. \end{aligned}$$

The relations (15.21) and (15.22), when inserted into (14.13), yield

$$\begin{aligned} \mathbf{c}_i(\tau_a) &= c_i, \\ \mathbf{c}_i(t) &= \begin{cases} \sum_j \bar{a}_{ij} \mathbf{c}''_j(t) & \text{if the root of } t \text{ is meagre,} \\ \sum_j a_{ij} \mathbf{c}'_j(t) & \text{if the root of } t \text{ is fat.} \end{cases} \end{aligned}$$

Finally, a comparison of the P-series for the exact and numerical solutions gives the order conditions (for order p)

$$\begin{aligned} \sum_i \bar{b}_i \mathbf{c}''_i(t) &= 1 \quad \text{for } t \in TN_q^a, \quad q = 2, \dots, p \\ \sum_i b_i \mathbf{c}'_i(t) &= 1 \quad \text{for } t \in TN_q^b, \quad q = 1, \dots, p. \end{aligned} \quad (15.23)$$

Exercises

1. Denote the number of elements of TP_q^a (P-trees with meagre root of order q) by α_q (see Table 15.2). Prove that

$$\alpha_1 + \alpha_2 x + \alpha_3 x^2 + \dots = (1-x)^{-2\alpha_1} (1-x^2)^{-2\alpha_2} (1-x^3)^{-2\alpha_3} \dots$$

Compute the first α_q and compare them with the a_q of Table 2.1.

Table 15.2. Number of elements of TP_q^a

q	1	2	3	4	5	6	7	8	9	10
α_q	1	2	7	26	107	458	2058	9498	44947	216598

2. There is no explicit, 4-stage Runge-Kutta method of order 4, which does not satisfy condition (1.9).

Hint. Use the techniques of the proof of Lemma 1.4.

3. Show that the order conditions (15.23) are the same as those given in Theorem 14.10.

4. Show that the partitioned method of Griepentrog (1978)

0	a_{ij}			0	0			\hat{a}_{ij}
1/2	1/2			1/2	$-\beta/2$	$(1+\beta)/2$		
1	-1	2		1	$(3+5\beta)/2$	$-(1+3\beta)$	$(1+\beta)/2$	
	1/6	2/3	1/6		1/6	2/3	1/6	

with $\beta = \sqrt{3}/3$ is of order 3 (the implicit method to the right is A -stable and is provided for the stiff part of the problem).

II.16 Symplectic Integration Methods

It is natural to look forward to those discrete systems which preserve as much as possible the intrinsic properties of the continuous system. (Feng Kang 1985)

Y.V. Rakitskii proposed . . . a requirement of the most complete conformity between two dynamical systems: one resulting from the original differential equations and the other resulting from the difference equations of the computational method. (Y.B. Suris 1989)

Hamiltonian systems, given by

$$\dot{p}_i = -\frac{\partial H}{\partial q_i}(p, q), \quad \dot{q}_i = \frac{\partial H}{\partial p_i}(p, q), \quad (16.1)$$

have been seen to possess two remarkable properties:

- a) the solutions preserve the Hamiltonian $H(p, q)$ (Ex. 5 of Section I.6);
- b) the corresponding flow is symplectic, i.e., preserves the differential 2-form

$$\omega^2 = \sum_{i=1}^n dp_i \wedge dq_i \quad (16.2)$$

(see Theorem I.14.12). In particular, the flow is volume preserving.

Both properties are usually destroyed by a numerical method applied to (16.1).

After some pioneering papers (de Vogelaere 1956, Ruth 1983, and Feng Kang (冯康) 1985) an enormous avalanche of research started around 1988 on the characterization of existing numerical methods which preserve symplecticity or on the construction of new classes of symplectic methods. An excellent overview is presented by Sanz-Serna (1992).

Example 16.1. We consider the harmonic oscillator

$$H(p, q) = \frac{1}{2} (p^2 + k^2 q^2). \quad (16.3)$$

Here (16.1) becomes

$$\dot{p} = -k^2 q, \quad \dot{q} = p \quad (16.4)$$

and we study the action of several steps of a numerical method on a well-known set of initial data (p_0, q_0) (see Fig. 16.1):

- a) The explicit Euler method (I.7.3)

$$\begin{pmatrix} p_m \\ q_m \end{pmatrix} = \begin{pmatrix} 1 & -hk^2 \\ h & 1 \end{pmatrix} \begin{pmatrix} p_{m-1} \\ q_{m-1} \end{pmatrix}, \quad h = \frac{\pi}{8k}, \quad m = 1, \dots, 16; \quad (16.5a)$$

b) the implicit (or backward) Euler method (7.3)

$$\begin{pmatrix} p_m \\ q_m \end{pmatrix} = \frac{1}{1 + h^2 k^2} \begin{pmatrix} 1 & -hk^2 \\ h & 1 \end{pmatrix} \begin{pmatrix} p_{m-1} \\ q_{m-1} \end{pmatrix}, \quad h = \frac{\pi}{8k}, \quad m = 1, \dots, 16; \quad (16.5b)$$

c) Runge's method (1.4) of order 2

$$\begin{pmatrix} p_m \\ q_m \end{pmatrix} = \begin{pmatrix} 1 - \frac{h^2 k^2}{2} & -hk^2 \\ h & 1 - \frac{h^2 k^2}{2} \end{pmatrix} \begin{pmatrix} p_{m-1} \\ q_{m-1} \end{pmatrix}, \quad h = \frac{\pi}{4k}, \quad m = 1, \dots, 8; \quad (16.5c)$$

d) the implicit midpoint rule (7.4) of order 2

$$\begin{pmatrix} p_m \\ q_m \end{pmatrix} = \frac{1}{1 + \frac{h^2 k^2}{4}} \begin{pmatrix} 1 - \frac{h^2 k^2}{4} & -hk^2 \\ h & 1 - \frac{h^2 k^2}{4} \end{pmatrix} \begin{pmatrix} p_{m-1} \\ q_{m-1} \end{pmatrix}, \quad h = \frac{\pi}{4k}, \quad m = 1, \dots, 8. \quad (16.5d)$$

For the exact flow, the last of all these cats would precisely coincide with the first one and all cats would have the same area. Only the last method appears to be area preserving. It also preserves the Hamiltonian in this example.

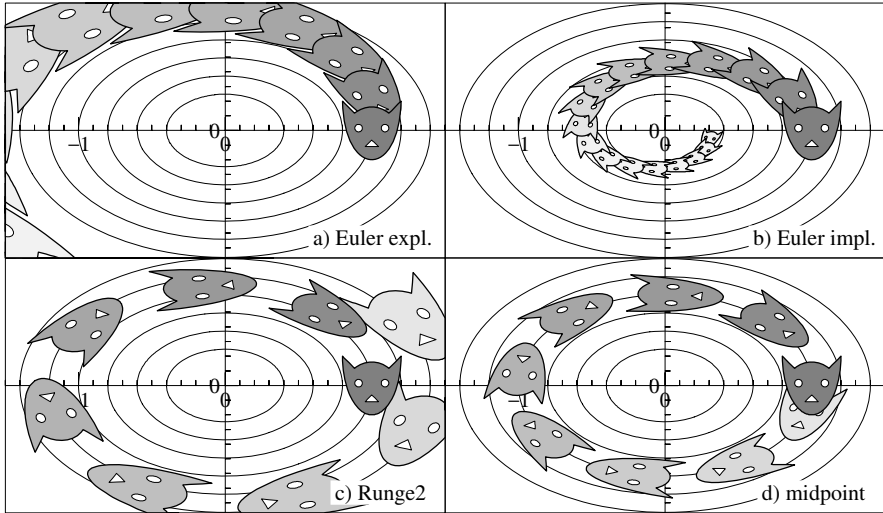


Fig. 16.1. Destruction of symplecticity of a Hamiltonian flow, $k = (\sqrt{5} + 1)/2$

Example 16.2. For a nonlinear problem we choose

$$H(p, q) = \frac{p^2}{2} - \cos(q) \left(1 - \frac{p}{6}\right) \quad (16.6)$$

which is similar to the Hamiltonian of the pendulum (I.14.25), but with some of the pendulum's symmetry destroyed. Fig. 16.2 presents 12000 consecutive solution values (p_i, q_i) for

- a) Runge's method of order 2 (see (1.4));
- b) the implicit Radau method with $s = 2$ and order 3 (see Exercise 6 of Section II.7);
- c) the implicit midpoint rule (7.4) of order 2.

The initial values are

$$p_0 = 0, \quad q_0 = \begin{cases} \arccos(0.5) = \pi/3 & \text{for case (a)} \\ \arccos(-0.8) & \text{for cases (b) and (c).} \end{cases}$$

The computation is done with fixed step sizes

$$h = \begin{cases} 0.15 & \text{for case (a)} \\ 0.3 & \text{for cases (b) and (c).} \end{cases}$$

The solution of method (a) spirals out, that of method (b) spirals in and both by no means preserve the Hamiltonian. Method (c) behaves differently. Although the Hamiltonian is not precisely preserved (see picture (d)), its error remains bounded for long-scale computations.

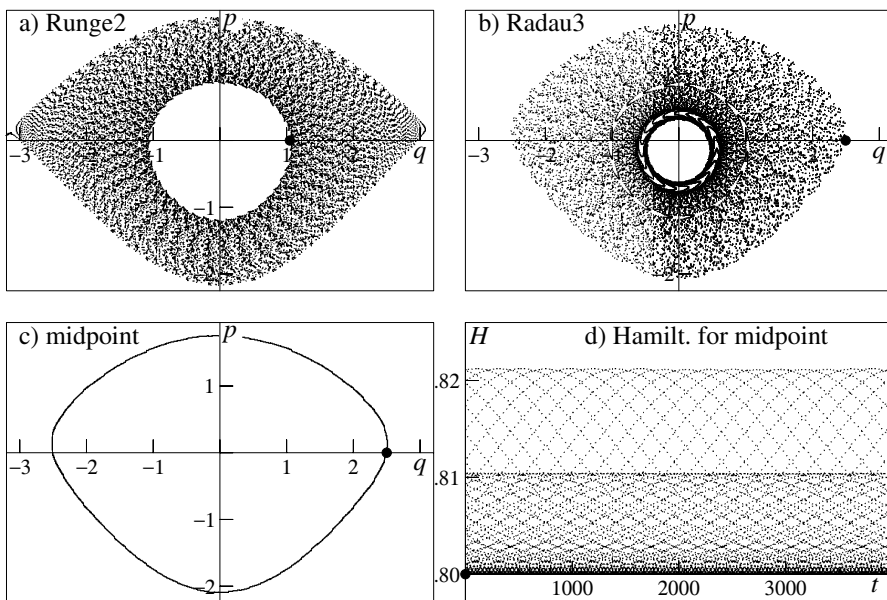


Fig. 16.2. A nonlinear pendulum and behaviour of H
 (• ... indicates the initial position)

Symplectic Runge-Kutta Methods

For a given Hamiltonian system (16.1), for a chosen one-step method (in particular a Runge-Kutta method) and a chosen step size h we denote by

$$\begin{aligned} \psi_h : \quad \mathbb{R}^{2n} &\longrightarrow \mathbb{R}^{2n} \\ (p_0, q_0) &\longmapsto (p_1, q_1) \end{aligned} \quad (16.7)$$

the transformation defined by the method.

Remark. For implicit methods the numerical solution (p_1, q_1) need not exist for all h and all initial values (p_0, q_0) nor need it be uniquely determined (see Exercise 2). Therefore we usually will have to restrict the domain where ψ_h is defined and we will have to select a solution of the nonlinear system such that ψ_h is differentiable on this domain. The subsequent results hold for all possible choices of ψ_h .

Definition 16.4. A one-step method is called *symplectic* if for every smooth Hamiltonian H and for every step size h the mapping ψ_h is symplectic (see Definition I.14.11), i.e., preserves the differential 2-form ω^2 of (16.2).

We start with the easiest result.

Theorem 16.5. *The implicit s -stage Gauss methods of order $2s$ (Kuntzmann & Butcher methods of Section II.7) are symplectic for all s .*

Proof. We simplify the notation by putting $h = 1$ and $t_0 = 0$ and use the fact that the methods under consideration are collocation methods, i.e., the numerical solution after one step is defined by $(u(1), v(1))$ where $(u(t), v(t))$ are polynomials of degree s such that

$$\begin{aligned} u(0) &= p_0, & u'(c_i) &= -\frac{\partial H}{\partial q}(u(c_i), v(c_i)) \\ v(0) &= q_0, & v'(c_i) &= \frac{\partial H}{\partial p}(u(c_i), v(c_i)) \end{aligned} \quad i = 1, \dots, s. \quad (16.8)$$

The polynomials $u(t)$ and $v(t)$ are now considered as functions of the initial values. For arbitrary variations ξ_1^0 and ξ_2^0 of the initial point we denote the corresponding variations of u and v as

$$\xi_1^t = \frac{\partial(u(t), v(t))}{\partial(p_0, q_0)} \cdot \xi_1^0, \quad \xi_2^t = \frac{\partial(u(t), v(t))}{\partial(p_0, q_0)} \cdot \xi_2^0.$$

Symplecticity of the method means that the expression

$$\omega^2(\xi_1^1, \xi_2^1) - \omega^2(\xi_1^0, \xi_2^0) = \int_0^1 \frac{d}{dt} \omega^2(\xi_1^t, \xi_2^t) dt \quad (16.9)$$

should vanish. Since ξ_1^t and ξ_2^t are polynomials in t of degree s , the expression $\frac{d}{dt} \omega^2(\xi_1^t, \xi_2^t)$ is a polynomial of degree $2s - 1$. We can thus exactly integrate (16.9)

by the Gaussian quadrature formula and so obtain

$$\omega^2(\xi_1^1, \xi_2^1) - \omega^2(\xi_1^0, \xi_2^0) = \sum_{i=1}^s b_i \frac{d}{dt} \omega^2(\xi_1^t, \xi_2^t) \Big|_{t=c_i}. \quad (16.9')$$

Differentiation of (16.8) with respect to (p_0, q_0) shows that (ξ_1^t, ξ_2^t) satisfies the variational equation (I.14.27) at the collocation points $t = c_i$, $i = 1, \dots, s$. Therefore, the computations of the proof of Theorem I.14.12 imply that

$$\frac{d}{dt} \omega^2(\xi_1^t, \xi_2^t) \Big|_{t=c_i} = 0 \quad \text{for } i = 1, \dots, s. \quad (16.10)$$

This, introduced into (16.9'), completes the proof of symplecticity. \square

The following theorem, discovered independently by at least three authors (F. Lasagni 1988, J.M. Sanz-Serna 1988, Y.B. Suris 1989) characterizes the class of all symplectic Runge-Kutta methods:

Theorem 16.6. *If the $s \times s$ matrix M with elements*

$$m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j, \quad i, j = 1, \dots, s \quad (16.11)$$

satisfies $M = 0$, then the Runge-Kutta method (7.7) is symplectic.

Proof. The matrix M has been known from nonlinear stability theory for many years (see Theorem IV.12.4). Both theorems have very similar proofs, the one works with the *inner* product, the other with the *exterior* product.

We write method (7.7) applied to problem (16.1) as

$$P_i = p_0 + h \sum_j a_{ij} k_j \quad Q_i = q_0 + h \sum_j a_{ij} \ell_j \quad (16.12a)$$

$$p_1 = p_0 + h \sum_i b_i k_i \quad q_1 = q_0 + h \sum_i b_i \ell_i \quad (16.12b)$$

$$k_i = -\frac{\partial H}{\partial q}(P_i, Q_i) \quad \ell_i = \frac{\partial H}{\partial p}(P_i, Q_i), \quad (16.12c)$$

denote the J th component of a vector by an upper index J and introduce the linear maps (one-forms)

$$\begin{aligned} dp_1^J : \mathbb{R}^{2n} &\rightarrow \mathbb{R}, & dP_i^J : \mathbb{R}^{2n} &\rightarrow \mathbb{R}, \\ \xi &\mapsto \frac{\partial p_1^J}{\partial (p_0, q_0)} \xi & \xi &\mapsto \frac{\partial P_i^J}{\partial (p_0, q_0)} \xi \end{aligned} \quad (16.13)$$

and similarly also dp_0^J , dk_i^J , dq_0^J , dq_1^J , dQ_i^J , $d\ell_i^J$ (the one-forms dp_0^J and dq_0^J correspond to dp_J and dq_J of Section I.14). Using the notation (16.13),

symplecticity of the method is equivalent to

$$\sum_{J=1}^n dp_1^J \wedge dq_1^J = \sum_{J=1}^n dp_0^J \wedge dq_0^J. \quad (16.14)$$

To check this relation we differentiate (16.12) with respect to the initial values and obtain

$$dP_i^J = dp_0^J + h \sum_j a_{ij} dk_j^J \quad dQ_i^J = dq_0^J + h \sum_j a_{ij} d\ell_j^J \quad (16.15a)$$

$$dp_1^J = dp_0^J + h \sum_i b_i dk_i^J \quad dq_1^J = dq_0^J + h \sum_i b_i d\ell_i^J \quad (16.15b)$$

$$dk_i^J = - \sum_{L=1}^n \frac{\partial^2 H}{\partial q^J \partial p^L}(P_i, Q_i) \cdot dP_i^L - \sum_{L=1}^n \frac{\partial^2 H}{\partial q^J \partial q^L}(P_i, Q_i) \cdot dQ_i^L \quad (16.15c)$$

$$d\ell_i^J = \sum_{L=1}^n \frac{\partial^2 H}{\partial p^J \partial p^L}(P_i, Q_i) \cdot dP_i^L + \sum_{L=1}^n \frac{\partial^2 H}{\partial p^J \partial q^L}(P_i, Q_i) \cdot dQ_i^L. \quad (16.15d)$$

We now compute

$$\begin{aligned} dp_1^J \wedge dq_1^J - dp_0^J \wedge dq_0^J \\ = h \sum_i b_i dp_0^J \wedge d\ell_i^J + h \sum_i b_i dk_i^J \wedge dq_0^J + h^2 \sum_{i,j} b_i b_j dk_i^J \wedge d\ell_j^J \end{aligned} \quad (16.16)$$

by using (16.15b) and the multilinearity of the wedge product. This formula corresponds precisely to (IV.12.6). Exactly as in the proof of Theorem IV.12.5, we now eliminate in (16.16) the quantities dp_0^J and dq_0^J with the help of (16.15a) to obtain

$$\begin{aligned} dp_1^J \wedge dq_1^J - dp_0^J \wedge dq_0^J \\ = h \sum_i b_i dP_i^J \wedge d\ell_i^J + h \sum_i b_i dk_i^J \wedge dQ_i^J - h^2 \sum_{i,j} m_{ij} dk_i^J \wedge d\ell_j^J, \end{aligned} \quad (16.17)$$

the formula analogous to (IV.12.7). Equations (16.15c,d) are perfect analogues of the variational equation (I.14.27). Therefore the same computations as in (I.14.39) give

$$\sum_{J=1}^n dP_i^J \wedge d\ell_i^J + \sum_{J=1}^n dk_i^J \wedge dQ_i^J = 0 \quad (16.18)$$

and the first two terms in (16.17) disappear. The last term vanishes by hypothesis (16.11) and we obtain (16.14). \square

Remark. F. Lasagni (1990) has proved in an unpublished manuscript that for *irreducible* methods (see Definitions IV.12.15 and IV.12.17) the condition $M = 0$ is also *necessary* for symplecticity. For a publication see Abia & Sanz-Serna (1993, Theorem 5.1), where this proof has been elaborated and adapted to a more general setting.

Remarks. a) Explicit Runge-Kutta methods are never symplectic (Ex. 1).

b) Equations (16.11) imply a substantial simplification of the order conditions (Sanz-Serna & Abia 1991). We shall return to this when treating partitioned methods (see (16.40)).

c) An important tool for the construction of symplectic methods is the W -transformation (see Section IV.5, especially Theorem IV.5.6). As can be seen from formula (IV.12.10), the method under consideration is symplectic if and only if the matrix X is skew-symmetric (with the exception of $x_{11} = 1/2$). Sun Geng (孙耿 1992) constructed several new classes of symplectic Runge-Kutta methods. One of his methods, based on Radau quadrature, is given in Table 16.1.

d) An inspection of Table IV.5.14 shows that all Radau IA, Radau IIA, Lobatto IIIA (in particular the trapezoidal rule), and Lobatto IIIC methods are not symplectic.

Table 16.1. Sun's symplectic Radau method of order 5

$\frac{4 - \sqrt{6}}{10}$	$\frac{16 - \sqrt{6}}{72}$	$\frac{328 - 167\sqrt{6}}{1800}$	$\frac{-2 + 3\sqrt{6}}{450}$
$\frac{4 + \sqrt{6}}{10}$	$\frac{328 + 167\sqrt{6}}{1800}$	$\frac{16 + \sqrt{6}}{72}$	$\frac{-2 - 3\sqrt{6}}{450}$
1	$\frac{85 - 10\sqrt{6}}{180}$	$\frac{85 + 10\sqrt{6}}{180}$	$\frac{1}{18}$
	$\frac{16 - \sqrt{6}}{36}$	$\frac{16 + \sqrt{6}}{36}$	$\frac{1}{9}$

Preservation of the Hamiltonian and of first integrals. In Exercise 5 of Section I.6 we have seen that the Hamiltonian $H(p, q)$ is a *first integral* of the system (16.1). This means that every solution $p(t), q(t)$ of (16.1) satisfies $H(p(t), q(t)) = \text{Const.}$ The numerical solution of a symplectic integrator does not share this property in general (see Fig. 16.2). However, we will show that every *quadratic* first integral will be preserved.

Denote $y = (p, q)$ and let G be a symmetric $2n \times 2n$ matrix. We suppose that the quadratic functional

$$\langle y, y \rangle_G := y^T G y$$

is a first integral of the system (16.1). This means that

$$\langle y, J^{-1} \text{grad } H(y) \rangle_G = 0 \quad \text{with} \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \quad (16.19)$$

for all $y \in \mathbb{R}^{2n}$.

Theorem 16.7 (Sanz-Serna 1988). *A symplectic Runge-Kutta method (i.e., a method satisfying (16.11)) leaves all quadratic first integrals of the system (16.1) invariant, i.e., the numerical solution $y_n = (p_n, q_n)$ satisfies*

$$\langle y_1, y_1 \rangle_G = \langle y_0, y_0 \rangle_G \quad (16.20)$$

for all symmetric matrices G satisfying (16.19).

Proof (Cooper 1987). The Runge-Kutta method (7.7) applied to problem (16.1) is given by

$$\begin{aligned} y_1 &= y_0 + \sum_i b_i k_i, & Y_i &= y_0 + \sum_j a_{ij} k_j, \\ k_i &= J^{-1} \text{grad } H(Y_i). \end{aligned} \quad (16.21)$$

As in the proof of Theorem 16.6 (see also Theorem IV.12.4) we obtain

$$\langle y_1, y_1 \rangle_G - \langle y_0, y_0 \rangle_G = 2h \sum_i b_i \langle Y_i, k_i \rangle_G - h^2 \sum_{i,j} m_{ij} \langle k_i, k_j \rangle_G.$$

The first term on the right-hand side vanishes by (16.19) and the second one by (16.11). \square

An Example from Galactic Dynamics

Always majestic, usually spectacularly beautiful, galaxies
are ... (Binney & Tremaine 1987)

While the theoretical meaning of symplecticity of numerical methods is clear, its importance for practical computations is less easy to understand. Numerous numerical experiments have shown that symplectic methods, in a fixed step size mode, show an excellent behaviour for long-scale scientific computations of Hamiltonian systems. We shall demonstrate this on the following example chosen from galactic dynamics and give a theoretical justification later in this section. However, Calvo & Sanz-Serna (1992c) have made the interesting discovery that *variable step size* implementation can *destroy* the advantages of symplectic methods. In order to illustrate this phenomenon we shall include in our computations violent step changes; one with a random number generator and one with the step size changing in function of the solution position.

A galaxy is a set of N stars which are mutually attracted by Newton's law. A relatively easy way to study them is to perform a long-scale computation of the orbit of *one* of its stars in the potential formed by the $N - 1$ remaining ones (see Binney & Tremaine 1987, Chapter 3); this potential is assumed to perform a uniform rotation with time, but not to change otherwise. The potential is determined

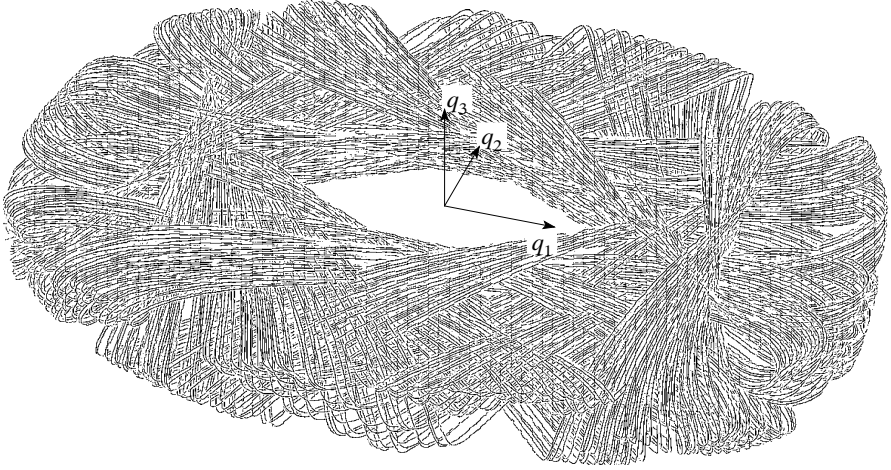


Fig. 16.3. Galactic orbit

by Poisson's differential equation $\Delta V = 4G\pi\rho$, where ρ is the density distribution of the galaxy, and real-life potential-density pairs are difficult to obtain (e.g., de Zeeuw & Pfenniger 1988). A popular issue is to choose a simple formula for V in such a way that the resulting ρ corresponds to a reasonable galaxy, for example (Binney 1981, Binney & Tremaine 1987, p. 45f, Pfenniger 1990)

$$V = A \ln \left(C + \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} \right). \quad (16.22)$$

The Lagrangian for a coordinate system rotating with angular velocity Ω becomes

$$\mathcal{L} = \frac{1}{2} \left((\dot{x} - \Omega y)^2 + (\dot{y} + \Omega x)^2 + \dot{z}^2 \right) - V(x, y, z). \quad (16.23)$$

This gives with the coordinates (see (I.6.23))

$$\begin{aligned} p_1 &= \frac{\partial \mathcal{L}}{\partial \dot{x}} = \dot{x} - \Omega y, & p_2 &= \frac{\partial \mathcal{L}}{\partial \dot{y}} = \dot{y} + \Omega x, & p_3 &= \frac{\partial \mathcal{L}}{\partial \dot{z}} = \dot{z}, \\ q_1 &= x, & q_2 &= y, & q_3 &= z, \end{aligned}$$

the Hamiltonian

$$\begin{aligned} H &= p_1 \dot{q}_1 + p_2 \dot{q}_2 + p_3 \dot{q}_3 - \mathcal{L} \\ &= \frac{1}{2} (p_1^2 + p_2^2 + p_3^2) + \Omega (p_1 q_2 - p_2 q_1) + A \ln \left(C + \frac{q_1^2}{a^2} + \frac{q_2^2}{b^2} + \frac{q_3^2}{c^2} \right). \end{aligned} \quad (16.24)$$

We choose the parameters and initial values as

$$\begin{aligned} a &= 1.25, \quad b = 1, \quad c = 0.75, \quad A = 1, \quad C = 1, \quad \Omega = 0.25, \\ q_1(0) &= 2.5, \quad q_2(0) = 0, \quad q_3(0) = 0, \quad p_1(0) = 0, \quad p_3(0) = 0.2, \end{aligned} \quad (16.25)$$

and take for $p_2(0)$ the larger of the roots for which $H = 2$. Our star then sets out for its voyage through the galaxy, the orbit is represented in Fig. 16.3 for $0 \leq t \leq 15000$. We are interested in its Poincaré sections with the half-plane $q_2 = 0$, $q_1 > 0$, $\dot{q}_2 > 0$ for $0 \leq t \leq 1000000$. These consist, for the exact solution, in 47101 cut points which are presented in Fig. 16.6l. These points were computed with the (non-symplectic) code DOP853 with $Tol = 10^{-17}$ in quadruple precision on a VAX 8700 computer.

Fig. 16.4, Fig. 16.5, and Fig. 16.6 present the obtained numerical results for the methods and step sizes summarized in Table 16.2.

Table 16.2. Methods for numerical experiments

item	method	order	h	points $t \leq 1000000$	impl.	symplec.	symmet.
a)	Gauss	6	1/5	47093	yes	yes	yes
b)	"	"	2/5	46852	"	"	"
c)	Gauss	6	random	46717	yes	yes	yes
d)	Gauss	6	partially halved	46576	yes	yes	yes
e)	Radau	5	1/10	46597	yes	no	no
f)	"	"	1/5	46266	"	"	"
g)	RK44	4	1/40	47004	no	no	no
h)	"	"	1/10	46192	"	"	"
i)	Lobatto	6	1/5	47091	yes	no	yes
j)	"	"	2/5	46839	"	"	"
k)	Sun Geng	5	1/5	47092	yes	yes	no
l)	exact	—	—	47101	—	—	—

Remarks.

- ad a): the Gauss6 method (Kuntzmann & Butcher method based on Gaussian quadrature with $s = 3$ and $p = 6$, see Table 7.4) for $h = 1/5$ is nearly identical to the exact solution;
- ad b): Gauss6 for $h = 2/5$ is much better than Gauss6 with random or partially halved step sizes (see item (c) and (d)) where $h \leq 2/5$.
- ad c): h was chosen at random uniformly distributed on $(0, 2/5)$;
- ad d): h was chosen “partially halved” in the sense that

$$h = \begin{cases} 2/5 & \text{if } q_1 > 0, \\ 1/5 & \text{if } q_1 < 0. \end{cases}$$

This produced the worst result for the 6th order Gauss method. We thus

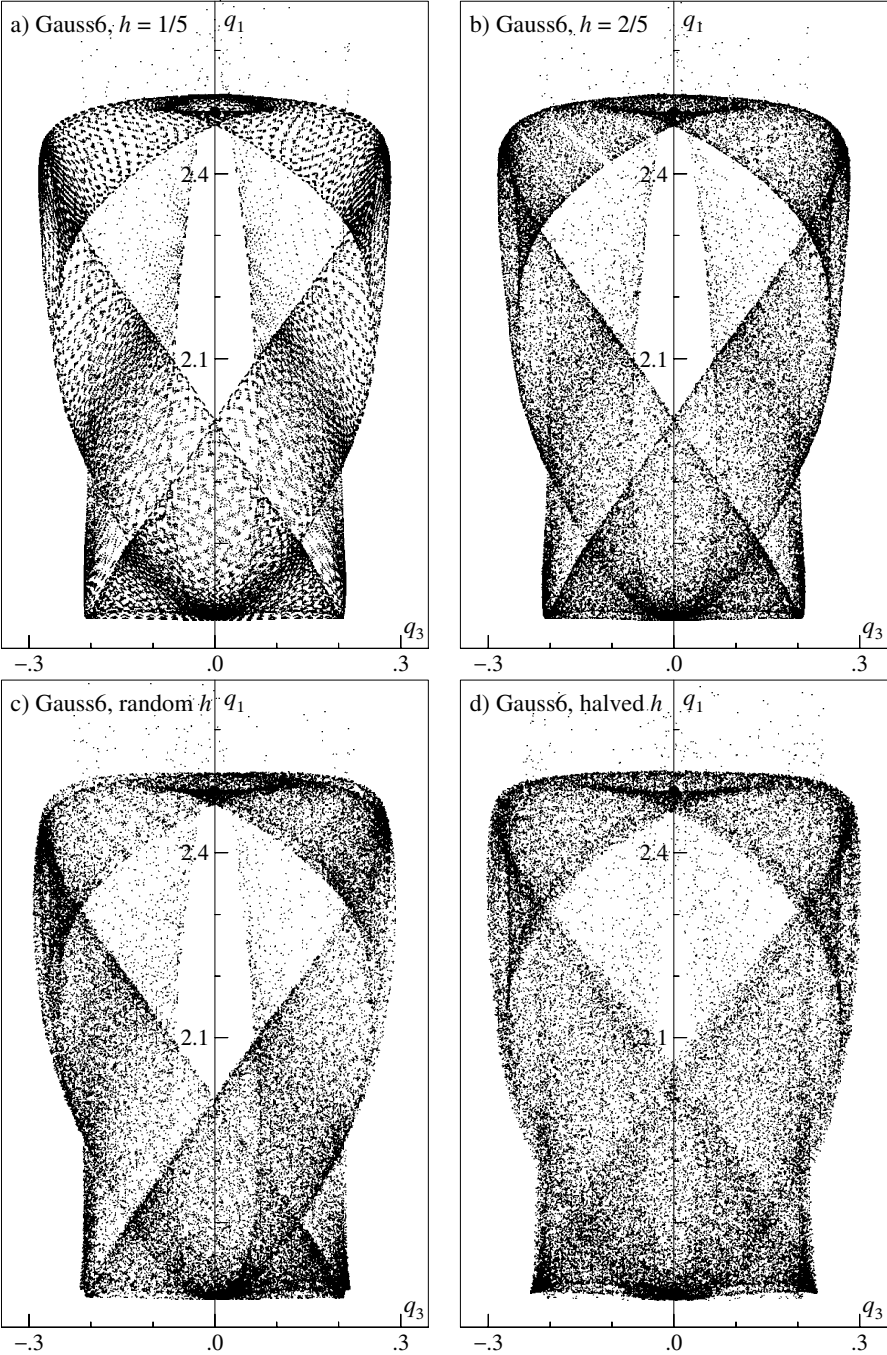


Fig. 16.4. Poincaré cuts for $0 \leq t \leq 1000000$; methods (a)-(d)

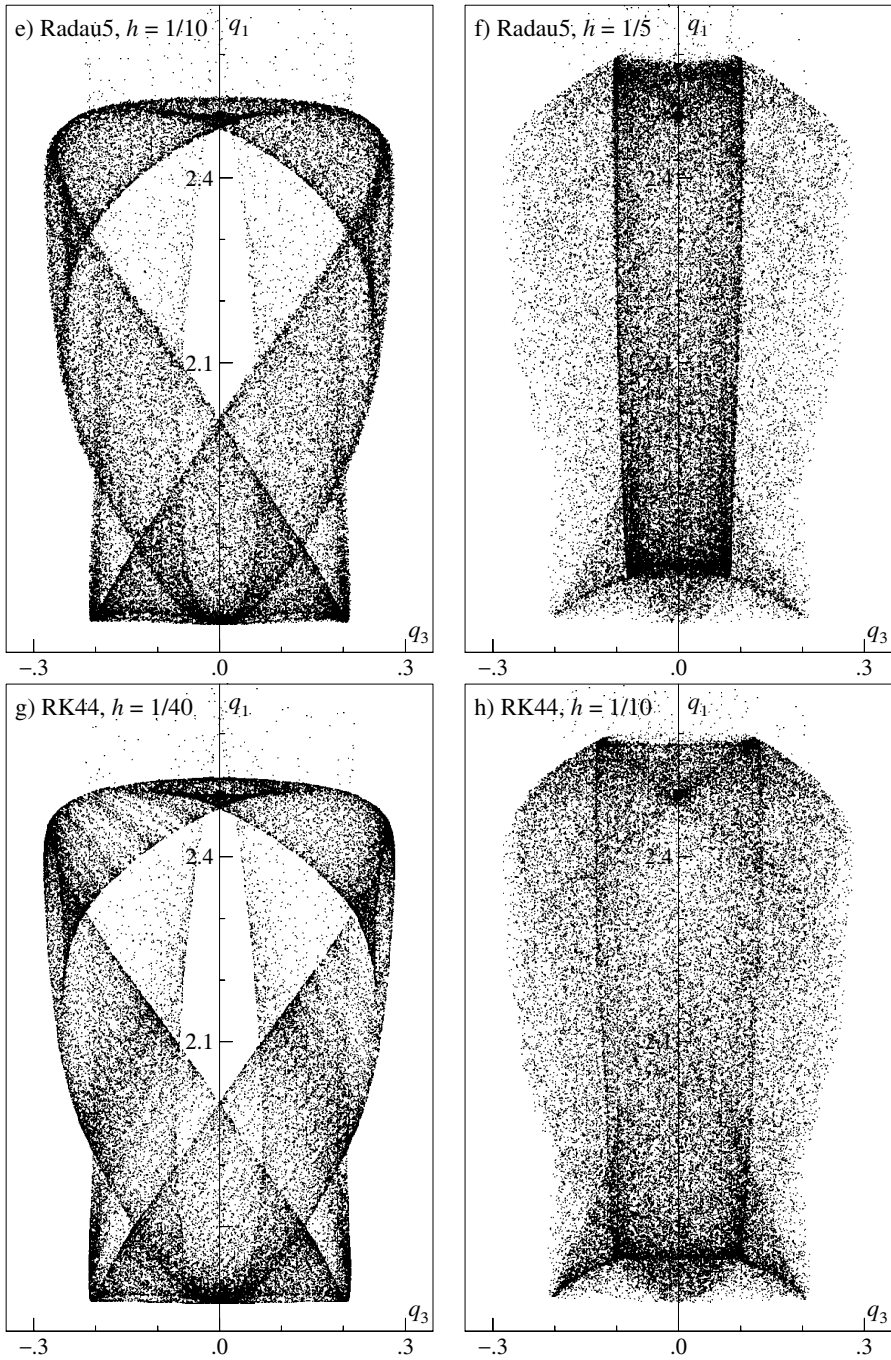


Fig. 16.5. Poincaré cuts for $0 \leq t \leq 1000000$; methods (e)-(h)

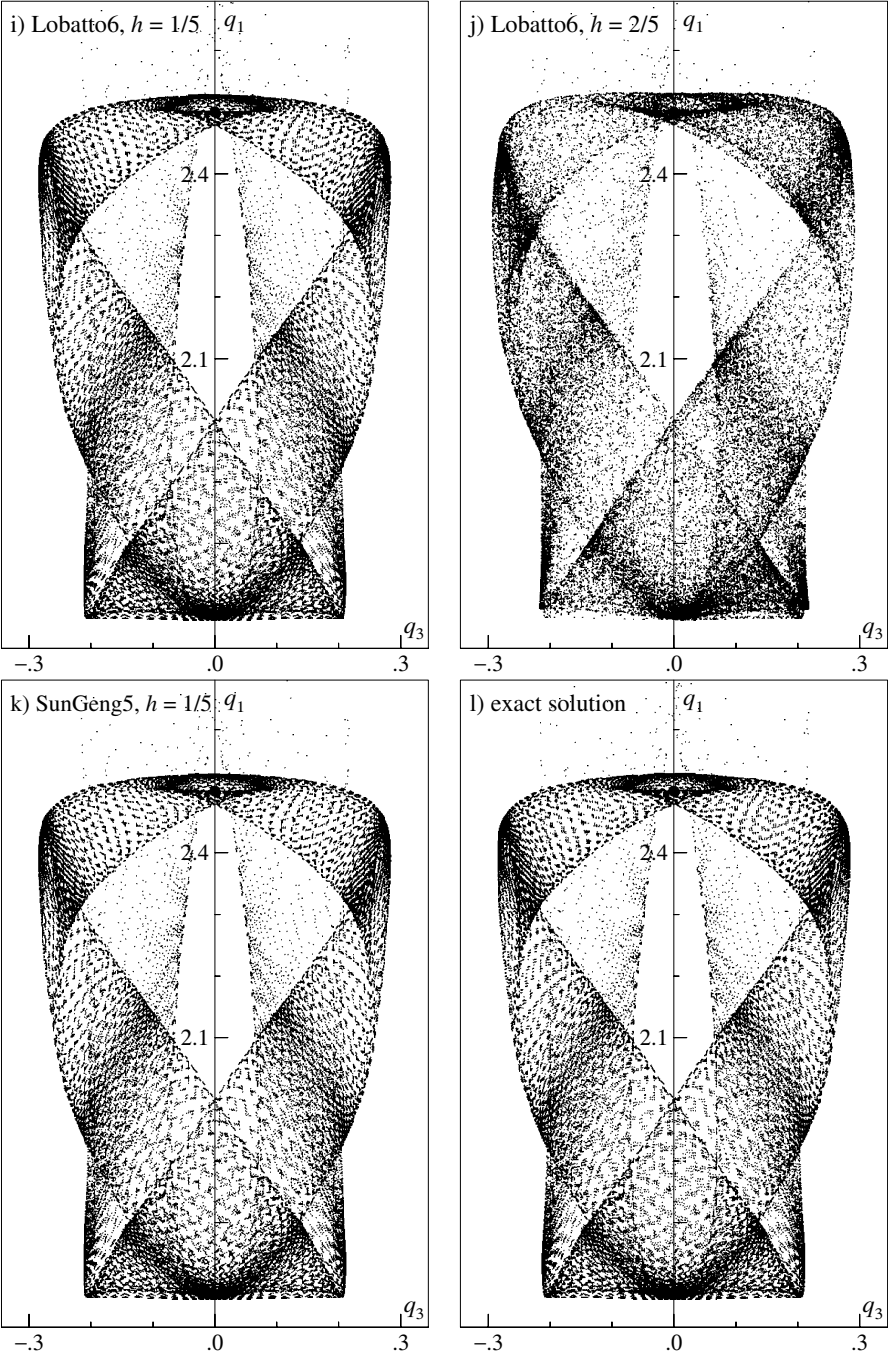


Fig. 16.6. Poincaré cuts for $0 \leq t \leq 1000000$; methods (i)-(l)

see that symplectic and symmetric methods compensate on the way back the errors committed on the outward journey.

- ad e), f): Radau5 (method of Ehle based on Radau quadrature with $s = 3$ and $p = 5$, see Table 7.7) is here not at all satisfactory;
- ad g): The explicit method RK44 (Runge-Kutta method with $s = p = 4$, see Table 1.2, left) is evidently much faster than the implicit methods, even with a smaller step size;
- ad h): With increasing step size RK44 deteriorates drastically;
- ad i): this is a non-symplectic but symmetric collocation method based on Lobatto quadrature with $s = 4$ of order 6 (see Table IV.5.8); its good performance on this nonlinear Hamiltonian problem is astonishing;
- ad j): with increasing h Lobatto6 is less satisfactory (see also Fig. 16.7);
- ad k): this is the symplectic non-symmetric method based on Radau quadrature of order 5 due to Sun Geng 孙耿 (Table 16.1).

The preservation of the Hamiltonian (correct value $H = 2$) during the computation for $0 \leq t \leq 1000000$ is shown in Fig. 16.7. While the errors for the symplectic and symmetric methods in constant step size mode remain bounded, random h (case c) results in a sort of Brownian motion, and the nonsymplectic methods as well as Gauss6 with partially halved step size result in permanent deterioration.

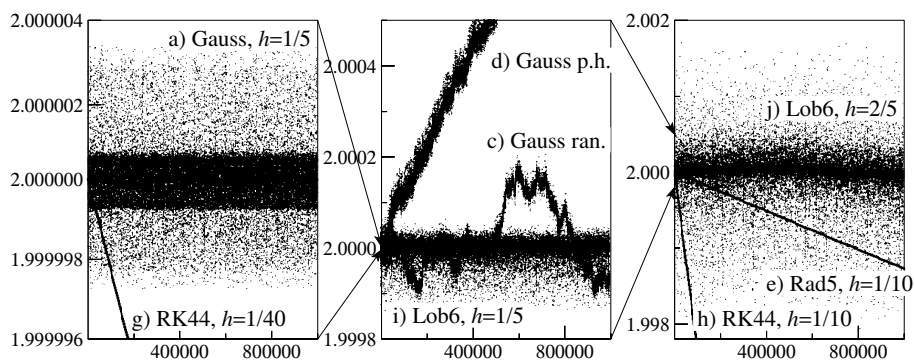


Fig. 16.7. Evolution of the Hamiltonian

Partitioned Runge-Kutta Methods

The fact that the system (16.1) possesses a natural partitioning suggests the use of partitioned Runge-Kutta methods as discussed in Section II.15. The main interest of such methods is for separable Hamiltonians where it is possible to obtain explicit symplectic methods.

A partitioned Runge-Kutta method for system (16.1) is defined by

$$P_i = p_0 + h \sum_j a_{ij} k_j \quad Q_i = q_0 + h \sum_j \hat{a}_{ij} \ell_j \quad (16.26a)$$

$$p_1 = p_0 + h \sum_i b_i k_i \quad q_1 = q_0 + h \sum_i \hat{b}_i \ell_i \quad (16.26b)$$

$$k_i = -\frac{\partial H}{\partial q}(P_i, Q_i) \quad \ell_i = \frac{\partial H}{\partial p}(P_i, Q_i) \quad (16.26c)$$

where b_i, a_{ij} and \hat{b}_i, \hat{a}_{ij} represent two different Runge-Kutta schemes.

Theorem 16.10 (Sanz-Serna 1992b, Suris 1990). *a) If the coefficients of (16.26) satisfy*

$$b_i = \hat{b}_i, \quad i = 1, \dots, s \quad (16.27)$$

$$b_i \hat{a}_{ij} + \hat{b}_j a_{ji} - b_i \hat{b}_j = 0, \quad i, j = 1, \dots, s \quad (16.28)$$

then the method (16.26) is symplectic.

b) If the Hamiltonian is separable (i.e., $H(p, q) = T(p) + U(q)$) then the condition (16.28) alone implies symplecticity of the method.

Proof. Following the lines of the proof of Theorem 16.6 we obtain

$$\begin{aligned} dp_1^J \wedge dq_1^J - dp_0^J \wedge dq_0^J &= h \sum_i \hat{b}_i dP_i^J \wedge d\ell_i^J + h \sum_i b_i dk_i^J \wedge dQ_i^J \\ &\quad - h^2 \sum_{i,j} (b_i \hat{a}_{ij} + \hat{b}_j a_{ji} - b_i \hat{b}_j) dk_i^J \wedge d\ell_j^J, \end{aligned} \quad (16.29)$$

instead of (16.17). The last term vanishes by (16.28). If $b_i = \hat{b}_i$ for all i , symplecticity of the method follows from (16.18). If the Hamiltonian is separable (the mixed derivatives $\partial^2 H / \partial q^J \partial p^L$ and $\partial^2 H / \partial p^J \partial q^L$ are not present in (16.15c,d)) then each of the two terms in (16.18) vanishes separately and the method is symplectic without imposing (16.27). \square

Remark. If (16.28) is satisfied and if the Hamiltonian is separable, it can be assumed without loss of generality that

$$b_i \neq 0, \quad \hat{b}_i \neq 0 \quad \text{for all } i. \quad (16.30)$$

Indeed, the stage values P_i (for i with $\widehat{b}_i = 0$) and Q_j (for j with $b_j = 0$) don't influence the numerical solution (p_1, q_1) and can be removed from the scheme. Notice however that in the resulting scheme the number of stages P_i may be different from that of Q_j .

Explicit methods for separable Hamiltonians. Let the Hamiltonian be of the form $H(p, q) = T(p) + U(q)$ and consider a partitioned Runge-Kutta method satisfying

$$\begin{aligned} a_{ij} &= 0 & \text{for } i < j & \quad (\text{diagonally implicit}) \\ \widehat{a}_{ij} &= 0 & \text{for } i \leq j & \quad (\text{explicit}). \end{aligned} \quad (16.31)$$

Since $\partial H / \partial q$ depends only on q , the method (16.26) is explicit for such a choice of coefficients. Under the assumption (16.30), the symplecticity condition (16.28) then becomes

$$a_{ij} = b_j \quad \text{for } i \geq j, \quad \widehat{a}_{ij} = \widehat{b}_j \quad \text{for } i > j, \quad (16.32)$$

so that the method (16.26) is characterized by the two schemes

$$\left| \begin{array}{cccccc} b_1 & & & & & \\ b_1 & b_2 & & & & \\ b_1 & b_2 & b_3 & & & \\ \vdots & \vdots & \ddots & \ddots & & \\ b_1 & b_2 & \cdots & b_{s-1} & b_s \\ \hline b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array} \right| \quad \left| \begin{array}{cccccc} 0 & & & & & \\ \widehat{b}_1 & 0 & & & & \\ \widehat{b}_1 & \widehat{b}_2 & 0 & & & \\ \vdots & \vdots & \ddots & \ddots & & \\ \widehat{b}_1 & \widehat{b}_2 & \cdots & \widehat{b}_{s-1} & 0 \\ \hline \widehat{b}_1 & \widehat{b}_2 & \cdots & \widehat{b}_{s-1} & \widehat{b}_s \end{array} \right| \quad (16.33)$$

If we admit the cases $b_1 = 0$ and/or $\widehat{b}_s = 0$, it can be shown (Exercise 6) that this scheme already represents the most general method (16.26) which is symplectic and explicit. We denote this scheme by

$$\begin{aligned} b: & \quad b_1 & b_2 & \cdots & b_s \\ \widehat{b}: & \quad \widehat{b}_1 & \widehat{b}_2 & \cdots & \widehat{b}_s. \end{aligned} \quad (16.34)$$

This method is particularly easy to implement:

$$\begin{aligned} P_0 &= p_0, \quad Q_1 = q_0 \\ \text{for } i &:= 1 \text{ to } s \text{ do} \\ P_i &= P_{i-1} - h b_i \partial U / \partial q(Q_i) \\ Q_{i+1} &= Q_i + h \widehat{b}_i \partial T / \partial p(P_i) \\ p_1 &= P_s, \quad q_1 = Q_{s+1} \end{aligned} \quad (16.35)$$

Special case $s = 1$. The combination of the implicit Euler method ($b_1 = 1$) with the explicit Euler method ($\widehat{b}_1 = 1$) gives the following symplectic method of order 1:

$$p_1 = p_0 - h \frac{\partial U}{\partial q}(q_0), \quad q_1 = q_0 + h \frac{\partial T}{\partial p}(p_1). \quad (16.36a)$$

By interchanging the roles of p and q we obtain the method

$$q_1 = q_0 + h \frac{\partial T}{\partial p}(p_0), \quad p_1 = p_0 - h \frac{\partial U}{\partial q}(q_1) \quad (16.36b)$$

which is also symplectic. Methods (16.36a) and (16.36b) are mutually adjoint (see Section II.8).

Construction of higher order methods. The order conditions for general partitioned Runge-Kutta methods applied to general problems (15.2) are derived in Section II.15 (Theorem 15.9). Let us here discuss how these conditions simplify in our special situation.

A) We consider the system (16.1) with separable Hamiltonian. In the notation of Section II.15 this means that $f_a(y_a, y_b)$ depends only on y_b and $f_b(y_a, y_b)$ depends only on y_a . Therefore, many elementary differentials vanish and only P-trees whose meagre and fat vertices alternate in each branch have to be considered. This is a considerable reduction of the order conditions.

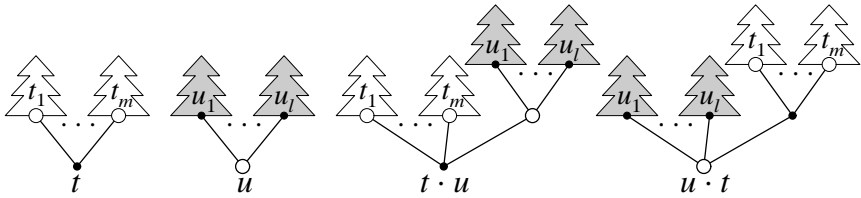


Fig. 16.8. Product of P-trees

B) As observed by Abia & Sanz-Serna (1993) the condition (16.28) acts as a simplifying assumption. Indeed, multiplying (16.28) by $\Phi_i(t) \cdot \Phi_j(u)$ (where $t = {}_a[t_1, \dots, t_m] \in TP^a$, $u = {}_b[u_1, \dots, u_l] \in TP^b$) and summing up over all i and j yields

$$\sum_i b_i \Phi_i(t \cdot u) + \sum_j \hat{b}_j \Phi_j(u \cdot t) - \left(\sum_i b_i \Phi_i(t) \right) \left(\sum_j \hat{b}_j \Phi_j(u) \right) = 0. \quad (16.37)$$

Here we have used the notation of Butcher (1987)

$$t \cdot u = {}_a[t_1, \dots, t_m, u], \quad u \cdot t = {}_b[u_1, \dots, u_l, t], \quad (16.38)$$

illustrated in Fig. 16.8. Since

$$\frac{1}{\gamma(t \cdot u)} + \frac{1}{\gamma(u \cdot t)} - \frac{1}{\gamma(t)} \cdot \frac{1}{\gamma(u)} = 0 \quad (16.39)$$

(this relation follows from (16.37) by inserting the coefficients of a symplectic Runge-Kutta method of sufficiently high order, e.g., a Gauss method) we obtain the following fact:

let $\varrho(t) + \varrho(u) = p$ and assume that all order conditions for P-trees of order $< p$ are satisfied, then

$$\sum_i b_i \Phi_i(t \cdot u) = \frac{1}{\gamma(t \cdot u)} \quad \text{iff} \quad \sum_j \hat{b}_j \Phi_j(u \cdot t) = \frac{1}{\gamma(u \cdot t)}. \quad (16.40)$$

From Fig. 16.8 we see that the P-trees $t \cdot u$ and $u \cdot t$ have the same geometrical structure. They differ only in the position of the root. Repeated application of this property implies that of all P-trees with identical geometrical structure only one has to be considered.

A method of order 3 (Ruth 1983). The above reductions leave five order conditions for a method of order 3 which, for $s = 3$, are the following:

$$\begin{aligned} b_1 + b_2 + b_3 &= 1, & \hat{b}_1 + \hat{b}_2 + \hat{b}_3 &= 1, & b_2 \hat{b}_1 + b_3(\hat{b}_1 + \hat{b}_2) &= 1/2, \\ b_2 \hat{b}_1^2 + b_3(\hat{b}_1 + \hat{b}_2)^2 &= 1/3, & \hat{b}_1 b_1^2 + \hat{b}_2(b_1 + b_2)^2 + \hat{b}_3(b_1 + b_2 + b_3)^2 &= 1/3. \end{aligned}$$

This nonlinear system possesses many solutions. A particularly simple solution, proposed by Ruth (1983), is

$$\begin{aligned} b: & \quad 7/24 \quad 3/4 \quad -1/24 \\ \hat{b}: & \quad 2/3 \quad -2/3 \quad 1. \end{aligned} \quad (16.41)$$

Concatenation of a method with its adjoint. The adjoint method of (16.26) is obtained by replacing h by $-h$ and by exchanging the roles of p_0, q_0 and p_1, q_1 (see Section II.8). This results in a partitioned Runge-Kutta method with coefficients (compare Theorem 8.3)

$$\begin{aligned} a_{ij}^* &= b_{s+1-j} - a_{s+1-i, s+1-j}, & b_i^* &= b_{s+1-i}, \\ \hat{a}_{ij}^* &= \hat{b}_{s+1-j} - \hat{a}_{s+1-i, s+1-j}, & \hat{b}_i^* &= \hat{b}_{s+1-i}. \end{aligned}$$

For the adjoint of (16.33) the first method is explicit and the second one is diagonally implicit, but otherwise it has the same structure. Adding dummy stages, it becomes of the form (16.33) with coefficients

$$\begin{aligned} b^*: & \quad 0 \quad b_s \quad b_{s-1} \quad \dots \quad b_1 \\ \hat{b}^*: & \quad \hat{b}_s \quad \hat{b}_{s-1} \quad \dots \quad \hat{b}_1 \quad 0. \end{aligned} \quad (16.42)$$

The following idea of Sanz-Serna (1992b) allows one to improve a method of odd order p : one considers the composition of method (16.33) (step size $h/2$) with its adjoint (again with step size $h/2$). The resulting method, which is represented by the coefficients

$$\begin{aligned} b_1/2 \quad b_2/2 \quad \dots \quad b_{s-1}/2 \quad b_s/2 \quad b_s/2 \quad b_{s-1}/2 \quad \dots \quad b_1/2 \\ \hat{b}_1/2 \quad \hat{b}_2/2 \quad \dots \quad \hat{b}_{s-1}/2 \quad \hat{b}_s \quad \hat{b}_{s-1}/2 \quad \dots \quad \hat{b}_1/2 \quad 0, \end{aligned}$$

is symmetric and therefore has an even order which is $\geq p + 1$. Concatenating

Ruth's method (16.41) with its adjoint yields the fourth order method

$$\begin{array}{cccccc} b: & 7/48 & 3/8 & -1/48 & -1/48 & 3/8 & 7/48 \\ \widehat{b}: & 1/3 & -1/3 & 1 & -1/3 & 1/3 & 0. \end{array} \quad (16.43)$$

Symplectic Nyström Methods

A frequent special case of a separable Hamiltonian $H(p, q) = T(p) + U(q)$ is when $T(p)$ is a quadratic functional $T(p) = p^T M p / 2$ (with M a constant symmetric matrix). In this situation the Hamiltonian system becomes

$$\dot{p} = -\frac{\partial U}{\partial q}(q), \quad \dot{q} = Mp,$$

which is equivalent to the second order equation

$$\ddot{q} = -M \frac{\partial U}{\partial q}(q). \quad (16.44)$$

It is therefore natural to consider Nyström methods (Section II.14) which for the system (16.44) are given by

$$\begin{aligned} Q_i &= q_0 + c_i h \dot{q}_0 + h^2 \sum_j \bar{a}_{ij} k'_j, & k'_j &= -M \frac{\partial U}{\partial q}(Q_j), \\ q_1 &= q_0 + h \dot{q}_0 + h^2 \sum_i \bar{b}_i k'_i, & \dot{q}_1 &= \dot{q}_0 + h \sum_i b_i k'_i. \end{aligned}$$

Replacing the variable \dot{q} by Mp and k'_i by $M\ell_i$, this method reads

$$\begin{aligned} Q_i &= q_0 + c_i h M p_0 + h^2 \sum_{j=1}^s \bar{a}_{ij} M \ell_j, & \ell_j &= -\frac{\partial U}{\partial q}(Q_j), \\ q_1 &= q_0 + h M p_0 + h^2 \sum_{i=1}^s \bar{b}_i M \ell_i, & p_1 &= p_0 + h \sum_{i=1}^s b_i \ell_i. \end{aligned} \quad (16.45)$$

Theorem 16.11 (Suris 1989). *Consider the system (16.44) where M is a symmetric matrix. Then, the s -stage Nyström method (16.45) is symplectic if the following two conditions are satisfied:*

$$\bar{b}_i = b_i(1 - c_i), \quad i = 1, \dots, s \quad (16.46a)$$

$$b_i(\bar{b}_j - \bar{a}_{ij}) = b_j(\bar{b}_i - \bar{a}_{ji}), \quad i, j = 1, \dots, s. \quad (16.46b)$$

Proof (Okunbor & Skeel 1992). As in the proof of Theorem 16.6 we differentiate the formulas (16.45) and compute

$$\begin{aligned} dp_1^J \wedge dq_1^J - dp_0^J \wedge dq_0^J \\ = h \sum_i b_i d\ell_i^J \wedge dq_0^J + h \sum_K M_{JK} dp_0^J \wedge dp_0^K \end{aligned} \quad (16.47)$$

$$\begin{aligned}
& + h^2 \sum_i b_i \sum_K M_{JK} d\ell_i^J \wedge dp_0^K + h^2 \sum_i \bar{b}_i \sum_K M_{JK} dp_0^J \wedge d\ell_i^K \\
& + h^3 \sum_{i,j} b_i \bar{b}_j \sum_K M_{JK} d\ell_i^J \wedge d\ell_j^K.
\end{aligned}$$

Next we eliminate dq_0^J with the help of the differentiated equation of Q_i , sum over all J and so obtain

$$\begin{aligned}
& \sum_{J=1}^n dp_1^J \wedge dq_1^J - \sum_{J=1}^n dp_0^J \wedge dq_0^J \\
& = h \sum_i b_i \sum_J d\ell_i^J \wedge dQ_i^J + h \sum_{J,K} M_{JK} dp_0^J \wedge dp_0^K \\
& + h^2 \sum_i (b_i - \bar{b}_i - b_i c_i) \sum_{J,K} M_{JK} d\ell_i^J \wedge dp_0^K \\
& + h^3 \sum_{i < j} (b_i \bar{b}_j - b_j \bar{b}_i - b_i \bar{a}_{ij} + b_j \bar{a}_{ji}) \sum_{J,K} M_{JK} d\ell_i^J \wedge d\ell_j^K.
\end{aligned}$$

The last two terms disappear by (16.46) whereas the first two terms vanish due to the symmetry of M and of the second derivatives of $U(q)$. \square

We have already encountered condition (16.46a) in Lemma 14.13. There, it was used as a simplifying assumption. It implies that only the order conditions for \dot{q}_1 have to be considered.

For Nyström methods satisfying both conditions of (16.46), one can assume without loss of generality that

$$b_i \neq 0 \quad \text{for } i = 1, \dots, s. \quad (16.48)$$

Let $I = \{i \mid b_i = 0\}$, then $\bar{b}_i = 0$ for $i \in I$ and $\bar{a}_{ij} = 0$ for $i \notin I, j \in I$. Hence, the stage values Q_i ($i \in I$) don't influence the numerical result (p_1, q_1) and can be removed from the scheme.

Explicit methods. Our main interest is in methods which satisfy

$$\bar{a}_{ij} = 0 \quad \text{for } i \leq j. \quad (16.49)$$

Under the assumption (16.48) the condition (16.46) then implies that the remaining coefficients are given by

$$\bar{a}_{ij} = b_j(c_i - c_j) \quad \text{for } i > j. \quad (16.50)$$

In this situation we may also suppose that

$$c_i \neq c_{i-1} \quad \text{for } i = 2, 3, \dots, s,$$

because equal consecutive c_i lead (via condition (16.50)) to equal stage values Q_i . Therefore the method is equivalent to one with a smaller number of stages.

The particular form of the coefficients \bar{a}_{ij} allows the following simple implementation (Okunbor & Skeel 1992b)

$$\begin{aligned}
 & Q_0 = q_0, \quad P_0 = p_0 \\
 & \text{for } i := 1 \text{ to } s \text{ do} \\
 & \quad Q_i = Q_{i-1} + h(c_i - c_{i-1})MP_{i-1} \quad (\text{with } c_0 = 0) \\
 & \quad P_i = P_{i-1} - hb_i \partial U / \partial q(Q_i) \\
 & \quad q_1 = Q_s + h(1 - c_s)MP_s, \quad p_1 = P_s.
 \end{aligned} \tag{16.51}$$

Special case $s = 1$. Putting $b_1 = 1$ (c_1 is a free parameter) yields a symplectic, explicit Nyström method of order 1. For the choice $c_1 = 1/2$ it has order 2.

Special case $s = 3$. To obtain order 3, four order conditions have to be satisfied (see Table 14.3). The first three mean that (b_i, c_i) is a quadrature formula of order 3. They allow us to express b_1, b_2, b_3 in terms of c_1, c_2, c_3 . The last condition then becomes (Okunbor & Skeel 1992b)

$$\begin{aligned}
 1 + 24\left(c_1 - \frac{1}{2}\right)\left(c_2 - \frac{1}{2}\right) + 24(c_2 - c_1)(c_3 - c_1)(c_3 - c_2) \\
 + 144\left(c_1 - \frac{1}{2}\right)\left(c_2 - \frac{1}{2}\right)\left(c_3 - \frac{1}{2}\right)\left(c_1 + c_3 - c_2 - \frac{1}{2}\right) = 0.
 \end{aligned} \tag{16.52}$$

We thus get a two-parameter family of third order methods. Okunbor & Skeel (1992b) suggest taking

$$c_2 = \frac{1}{2}, \quad c_1 = 1 - c_3 = \frac{1}{6}\left(2 + \sqrt[3]{2} + \frac{1}{\sqrt[3]{2}}\right) \tag{16.53}$$

(the real root of $12c_1(2c_1 - 1)^2 = 1$). This method is symmetric and thus of order 4. Another 3-stage method of order 4 has been found by Qin Meng-Zhao & Zhu Wen-jie (1991).

Higher order methods. For the construction of methods of order ≥ 4 it is worthwhile to investigate the effect of the condition (16.46b) on the order conditions. As for partitioned Runge-Kutta methods one can show that SN-trees with the same geometrical structure lead to equivalent order conditions. For details we refer to Calvo & Sanz-Serna (1992). With the notation of Table 14.3, the SN-trees t_6 and t_7 as well as the pairs t_9, t_{12} and t_{10}, t_{13} give rise to equivalent order conditions. Consequently, for order 5, one has to consider 10 conditions. Okunbor & Skeel (1992c) present explicit, symplectic Nyström methods of orders 5 and 6 with 5 and 7 stages, respectively. A 7th order method is given by Calvo & Sanz-Serna (1992b).

Conservation of the Hamiltonian; Backward Analysis

The differential equation actually solved by the difference scheme will be called the modified equation.

(Warming & Hyett 1974, p. 161)

The *wrong* solution of the *right* equation; the *right* solution of the *wrong* equation.

(Feng Kang, Beijing Sept. 1, 1992)

We have observed above (Example 16.2 and Fig. 16.6) that for the numerical solution of symplectic methods the Hamiltonian H remained between fixed bounds over any long-term integration, i.e., so-called secular changes of H were absent. Following several authors (Yoshida 1993, Sanz-Serna 1992, Feng Kang 1991b) this phenomenon is explained by interpreting the numerical solution as the *exact* solution of a *perturbed Hamiltonian system*, which is obtained as the formal expansion (16.56) in powers of h . The *exact* conservation of the perturbed Hamiltonian \tilde{H} then involves the quasi-periodic behaviour of H along the computed points. This resembles Wilkinson's famous idea of backward error analysis in linear algebra and, in the case of differential equations, seems to go back to Warming & Hyett (1974). We demonstrate this idea for the symplectic Euler method (see (16.36b))

$$\begin{aligned} p_1 &= p_0 - hH_q(p_0, q_1) \\ q_1 &= q_0 + hH_p(p_0, q_1) \end{aligned} \quad (16.54)$$

which, when expanded around the point (p_0, q_0) , gives

$$\begin{aligned} p_1 &= p_0 - hH_q - h^2 H_{qq} H_p - \frac{h^3}{2} H_{qqq} H_p H_p - h^3 H_{qq} H_{pq} H_p - \dots \Big|_{p_0, q_0} \\ q_1 &= q_0 + hH_p + h^2 H_{pq} H_p + \frac{h^3}{2} H_{pqq} H_p H_p + h^3 H_{pq} H_{pq} H_p + \dots \Big|_{p_0, q_0}. \end{aligned} \quad (16.54')$$

In the case of non-scalar equations the p 's and q 's must here be equipped with various summation indices. We suppress these in the sequel for the sake of simplicity and think of scalar systems only. The exact solution of a perturbed Hamiltonian

$$\begin{aligned} \dot{p} &= -\tilde{H}_q(p, q) \\ \dot{q} &= \tilde{H}_p(p, q) \end{aligned}$$

has a Taylor expansion analogous to Theorem 2.6 as follows

$$\begin{aligned} p_1 &= p_0 - h\tilde{H}_q + \frac{h^2}{2} (\tilde{H}_{qp}\tilde{H}_q - \tilde{H}_{qq}\tilde{H}_p) + \dots \\ q_1 &= q_0 + h\tilde{H}_p + \frac{h^2}{2} (-\tilde{H}_{pp}\tilde{H}_q + \tilde{H}_{pq}\tilde{H}_p) + \dots \end{aligned} \quad (16.55)$$

We now set

$$\tilde{H} = H + hH^{(1)} + h^2H^{(2)} + h^3H^{(3)} + \dots \quad (16.56)$$

with unknown functions $H^{(1)}, H^{(2)}, \dots$, insert this into (16.55) and compare the

resulting formulas with (16.54'). Then the comparison of the h^2 terms gives

$$H_q^{(1)} = \frac{1}{2} H_{qq} H_p + \frac{1}{2} H_{qp} H_q, \quad H_p^{(1)} = \frac{1}{2} H_{pp} H_q + \frac{1}{2} H_{pq} H_p$$

which by miracle (the "miracle" is in fact a consequence of the symplecticity of method (16.54)) allow the common primitive

$$H^{(1)} = \frac{1}{2} H_p H_q. \quad (16.56;1)$$

The h^3 terms lead to

$$H^{(2)} = \frac{1}{12} (H_{pp} H_q^2 + H_{qq} H_p^2 + 4H_{pq} H_p H_q) \quad (16.56;2)$$

and so on.

Connection with the Campbell-Baker-Hausdorff formula. An elegant access to the expansion (16.56), which works for separable Hamiltonians $H(p, q) = T(p) + U(q)$, has been given by Yoshida (1993). We interpret method (16.54) as composition of the two symplectic maps

$$z_0 = \begin{pmatrix} p_0 \\ q_0 \end{pmatrix} \xrightarrow{S_T} z = \begin{pmatrix} p_0 \\ q_1 \end{pmatrix} \xrightarrow{S_U} z_1 = \begin{pmatrix} p_1 \\ q_1 \end{pmatrix} \quad (16.57)$$

which consist, respectively, in solving exactly the Hamiltonian systems

$$\begin{aligned} \dot{p} &= 0 & \text{and} & & \dot{p} &= -U_q(q) \\ \dot{q} &= T_p(p) & & & \dot{q} &= 0 \end{aligned} \quad (16.58)$$

and apply some Lie theory. If we introduce for these equations the differential operators given by (13.2')

$$D_T \Psi = \frac{\partial \Psi}{\partial q} T_p(p), \quad D_U \Psi = -\frac{\partial \Psi}{\partial p} U_q(q), \quad (16.59)$$

the formulas (13.3) allow us to write the Taylor series of the map S_T as

$$z = \sum_{i=0}^{\infty} \frac{h^i}{i!} D_T^i z \Big|_{z=z_0}. \quad (16.60)$$

If now $F(z)$ is an arbitrary function of the solution $z(t) = (p(t), q(t))$ (left equation of (16.58)), we find, as in (13.2), that

$$F(z)' = D_T F, \quad F(z)'' = D_T^2 F, \dots$$

and (16.60) extends to (Gröbner 1960)

$$F(z) = \sum_{i=0}^{\infty} \frac{h^i}{i!} D_T^i F(z) \Big|_{z=z_0}. \quad (16.60')$$

We now insert S_U for F and insert for S_U the formula analogous to (16.60) to

obtain for the composition (16.57)

$$\begin{aligned} z_1 = (p_1, q_1) &= \sum_{i=0}^{\infty} \frac{h^i}{i!} D_T^i \sum_{j=0}^{\infty} \frac{h^j}{j!} D_U^j z \Big|_{z=z_0} \\ &= \exp(hD_T) \exp(hD_U)(p, q) \Big|_{p=p_0, q=q_0}. \end{aligned} \quad (16.61)$$

But the product $\exp(hD_T) \exp(hD_U)$ is *not* $\exp(hD_T + hD_U)$, as we have all learned in school, because the operators D_T and D_U do not commute. This is precisely the content of the famous Campbell-Baker-Hausdorff Formula (claimed in 1898 by J.E. Campbell and proved independently by Baker (1905) and in the “kleine Untersuchung” of Hausdorff (1906)) which states, for our problem, that

$$\exp(hD_T) \exp(hD_U) = \exp(h\tilde{D}) \quad (16.62)$$

where

$$\begin{aligned} \tilde{D} &= D_T + D_U + \frac{h}{2} [D_T, D_U] + \frac{h^2}{12} ([D_T, [D_T, D_U]] + [D_U, [D_U, D_T]]) \\ &\quad + \frac{h^3}{24} [D_T, [D_U, [D_U, D_T]]] + \dots \end{aligned} \quad (16.63)$$

and $[D_A, D_B] = D_A D_B - D_B D_A$ is the commutator. Equation (16.62) shows that the map (16.57) is the exact solution of the differential equation corresponding to the differential operator \tilde{D} . A straightforward calculation now shows: If

$$D_A \Psi = -\frac{\partial \Psi}{\partial p} A_q + \frac{\partial \Psi}{\partial q} A_p \quad \text{and} \quad D_B \Psi = -\frac{\partial \Psi}{\partial p} B_q + \frac{\partial \Psi}{\partial q} B_p \quad (16.64)$$

are differential operators corresponding to Hamiltonians A and B respectively, then

$$[D_A, D_B] \Psi = D_C \Psi = -\frac{\partial \Psi}{\partial p} C_q + \frac{\partial \Psi}{\partial q} C_p$$

where

$$C = A_p B_q - A_q B_p. \quad (16.65)$$

A repeated application of (16.65) now allows us to obtain for all brackets in (16.63) a corresponding Hamiltonian which finally leads to

$$\tilde{H} = T + U + \frac{h}{2} T_p U_q + \frac{h^2}{12} (T_{pp} U_q^2 + U_{qq} T_p^2) + \frac{h^3}{12} T_{pp} U_{qq} T_p U_q + \dots \quad (16.66)$$

which is the specialization of (16.56) to separable Hamiltonians.

Example 16.12 (Yoshida 1993). For the mathematical pendulum

$$H(p, q) = \frac{p^2}{2} - \cos q \quad (16.67)$$

series (16.66) becomes

$$\tilde{H} = \frac{p^2}{2} - \cos q + \frac{h}{2} p \sin q + \frac{h^2}{12} (\sin^2 q + p^2 \cos q) + \frac{h^3}{12} p \cos q \sin q + \mathcal{O}(h^4). \quad (16.68)$$

Fig. 16.9 presents for various step sizes h and for various initial points ($p_0=0, q_0=-1.5$; $p_0=0, q_0=-2.5$; $p_0=1.5, q_0=-\pi$; $p_0=2.5, q_0=-\pi$) the numerically computed points for method (16.54) compared to the contour lines of $\tilde{H} = \text{Const}$ given by the terms up to order h^3 in (16.68). The excellent agreement of the results with theory for $h \leq 0.6$ leaves nothing to be desired, while for h beyond 0.9 the dynamics of the numerical method turns rapidly into chaotic behaviour.

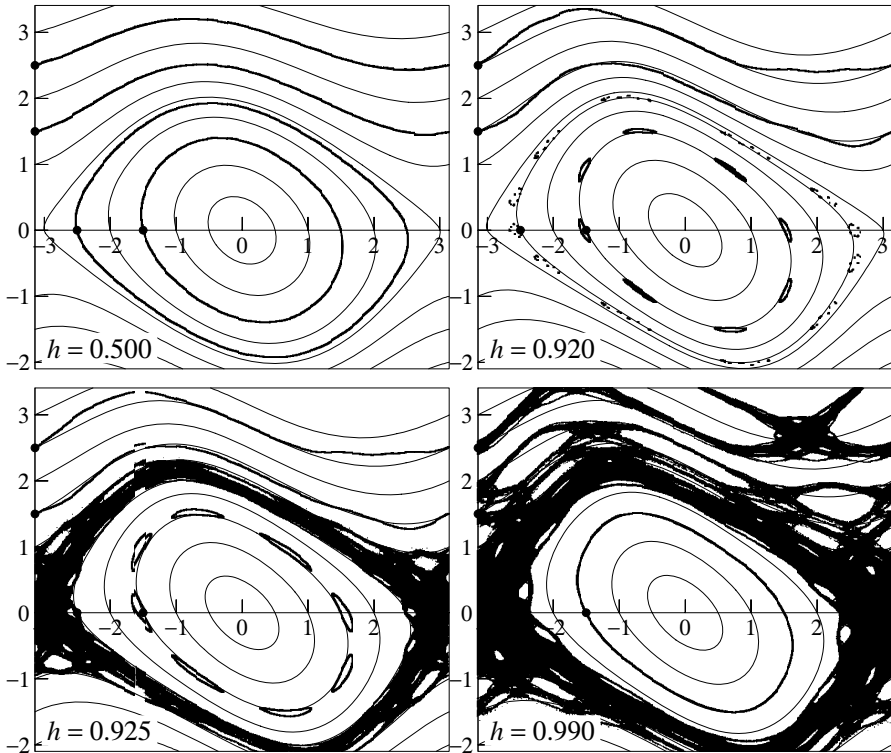


Fig. 16.9. Symplectic method compared to perturbed Hamiltonian
($\bullet \dots$ indicate the initial positions)

Remark. For much research, especially in the beginning of the “symplectic era”, the central role for the construction of canonical difference schemes is played by the Hamilton-Jacobi theory and generating functions. For this, the reader may consult the papers Feng Kang (1986), Feng Kang, Wu Hua-mo, Qin Meng-zhao & Wang Dao-liu (1989), Channell & Scovel (1990) and Miesbach & Pesch (1992). Many additional numerical experiments can be found in Channell & Scovel (1990), Feng Kang (1991), and Pullin & Saffman (1991).

Exercises

1. Show that explicit Runge-Kutta methods are never symplectic.

Hint. Compute the diagonal elements of M .

2. Study the existence and uniqueness of the numerical solution for the implicit mid-point rule when applied to the Hamiltonian system

$$\dot{p} = -q^2, \quad \dot{q} = p.$$

Show that the method possesses no solution at all for $h^2 q_0 + h^3 p_0/2 < -1$ and two solutions for $h^2 q_0 + h^3 p_0/2 > -1$ ($h \neq 0$). Only one of the solutions tends to (p_0, q_0) for $h \rightarrow 0$.

3. A Runge-Kutta method is called *linearly symplectic* if it is symplectic for all linear Hamiltonian systems

$$\dot{y} = J^{-1} C y$$

(J is given in (16.19) and C is a symmetric matrix). Prove (Feng Kang 1985) that a Runge-Kutta method is linearly symplectic if and only if its stability function satisfies

$$R(-z)R(z) = 1 \quad \text{for all } z \in \mathbb{C}. \quad (16.69)$$

Hint. For the definition of the stability function see Section IV.2 of Volume II. Then by Theorem I.14.14, linear symplecticity is equivalent to

$$R(hJ^{-1}C)^T J R(hJ^{-1}C) = J.$$

Furthermore, the matrix $B := J^{-1}C$ is seen to verify $B^T J = -JB$ and hence also $(B^k)^T J = J(-B)^k$ for $k = 0, 1, 2, \dots$. This implies that

$$R(hJ^{-1}C)^T J = J R(-hJ^{-1}C).$$

4. Prove that the stability function of a symmetric Runge-Kutta method satisfies (16.69).
5. Compute all quadratic first integrals of the Hamiltonian system (16.4).
6. For a separable Hamiltonian consider the method (16.26) where $a_{ij} = 0$ for $i < j$, $\hat{a}_{ij} = 0$ for $i < j$ and for every i either $a_{ii} = 0$ or $\hat{a}_{ii} = 0$. If the method satisfies (16.28) then it is equivalent to one given by scheme (16.33).

Hint. Remove first all stages which don't influence the numerical result (see the remark after Theorem 16.10). Then deduce from (16.28) relations similar to (16.32). Finally, remove identical stages and add, if necessary, a dummy stage in order that both methods have the same number of stages.

7. (Lasagni 1990). Characterize symplecticity for multi-derivative Runge-Kutta methods. Show that the s -stage q -derivative method of Definition 13.1 is symplectic if its coefficients satisfy

$$b_i^{(r)} b_j^{(m)} - b_i^{(r)} a_{ij}^{(m)} - b_j^{(m)} a_{ji}^{(r)} = \begin{cases} b_i^{(r+m)} & \text{if } i = j \text{ and } r + m \leq q, \\ 0 & \text{otherwise.} \end{cases} \quad (16.70)$$

Hint. Denote $k^{(r)} = D_H^r p$, $\ell^{(r)} = D_H^r q$, where D_H is the differential operator as in (16.59) and (16.64), so that the exact solution of (16.1) is given by

$$p(x_0+h) = p_0 + \sum_{r \geq 1} \frac{h^r}{r!} k^{(r)}(p_0, q_0), \quad q(x_0+h) = q_0 + \sum_{r \geq 1} \frac{h^r}{r!} \ell^{(r)}(p_0, q_0).$$

Then deduce from the symplecticity of the exact solution that

$$\frac{1}{\varrho!} (dp \wedge d\ell^{(\varrho)} + dk^{(\varrho)} \wedge dq) + \sum_{r+m=\varrho} \frac{1}{r!} \frac{1}{m!} dk^{(r)} \wedge d\ell^{(m)} = 0. \quad (16.71)$$

This, together with a modification of the proof of Theorem 16.6, allows us to obtain the desired result.

8. (Yoshida 1990, Qin Meng-Zhao & Zhu Wen-Jie 1992). Let $y_1 = \psi_h(y_0)$ denote a symmetric numerical scheme of order $p = 2k$. Prove that the composed method

$$\psi_{c_1 h} \circ \psi_{c_2 h} \circ \psi_{c_1 h}$$

is symmetric and has order $p + 2$ if

$$2c_1 + c_2 = 1, \quad 2c_1^{2k+1} + c_2^{2k+1} = 0. \quad (16.72)$$

Hence there exist, for separable Hamiltonians, explicit symplectic partitioned methods of arbitrarily high order.

Hint. Proceed as for (4.1)-(4.2) and use Theorem 8.10 (the order of a symmetric method is even).

9. The Hamiltonian function (16.24) for the galactic problem is *not* separable. Nevertheless, both methods (16.36a) and (16.36b) can be applied explicitly. Explain.

II.17 Delay Differential Equations

Detailed studies of the real world impel us, albeit reluctantly, to take account of the fact that the rate of change of physical systems depends not only on their present state, but also on their past history. (Bellman & Cooke 1963)

Delay differential equations are equations with “retarded arguments” or “time lags” such as

$$y'(x) = f(x, y(x), y(x - \tau)) \quad (17.1)$$

or

$$y'(x) = f(x, y(x), y(x - \tau_1), y(x - \tau_2)) \quad (17.2)$$

or of even more general form. Here the derivative of the solutions depends also on its values at previous points.

Time lags are present in many models of applied mathematics. They can also be the source of interesting mathematical phenomena such as instabilities, limit cycles, periodic behaviour.

Existence

For equations of the type (17.1) or (17.2), where the delay values $x - \tau$ are bounded away from x by a positive constant, the question of existence is an easy matter: suppose that the solution is known, say

$$y(x) = \varphi(x) \quad \text{for } x_0 - \tau \leq x \leq x_0.$$

Then $y(x - \tau)$ is a known function of x for $x_0 \leq x \leq x_0 + \tau$ and (17.1) becomes an ordinary differential equation, which can be treated by known existence theories. We then know $y(x)$ for $x_0 \leq x \leq x_0 + \tau$ and can compute the solution for $x_0 + \tau \leq x \leq x_0 + 2\tau$ and so on. This “method of steps” then yields existence and uniqueness results for all x . For more details we recommend the books of Bellman & Cooke (1963) and Driver (1977, especially Chapter V).

Example 1. We consider the equation

$$y'(x) = -y(x - 1), \quad y(x) = 1 \quad \text{for } -1 \leq x \leq 0. \quad (17.3)$$

Proceeding as described above, we obtain

$$\begin{aligned}
 y(x) &= 1 - x && \text{for } 0 \leq x \leq 1, \\
 y(x) &= 1 - x + \frac{(x-1)^2}{2!} && \text{for } 1 \leq x \leq 2, \\
 y(x) &= 1 - x + \frac{(x-1)^2}{2!} - \frac{(x-2)^3}{3!} && \text{for } 2 \leq x \leq 3, \text{ etc.}
 \end{aligned}$$

The solution is displayed in Fig. 17.1. We observe that despite the fact that the differential equation and the initial function are C^∞ , the solution has discontinuities in its derivatives. This results from the fact that the initial function does not satisfy the differential equation. With every time step τ , however, these discontinuities are smoothed out more and more.

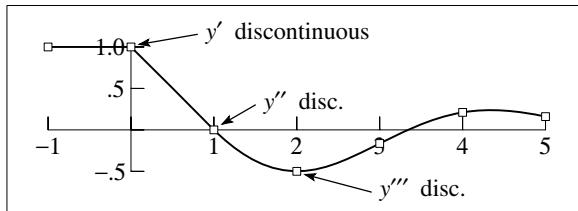


Fig. 17.1. Solution of (17.3)

Example 2. Our next example clearly illustrates the fact that the solutions of a delay equation depend on the entire history between $x_0 - \tau$ and x_0 , and not only on the initial value:

$$y'(x) = -1.4 \cdot y(x-1) \quad (17.4)$$

- a) $\varphi(x) = 0.8$ for $-1 \leq x \leq 0$,
- b) $\varphi(x) = 0.8 + x$ for $-1 \leq x \leq 0$,
- c) $\varphi(x) = 0.8 + 2x$ for $-1 \leq x \leq 0$.

The solutions are displayed in Fig. 17.2. An explanation for the oscillatory behaviour of the solutions will be given below.

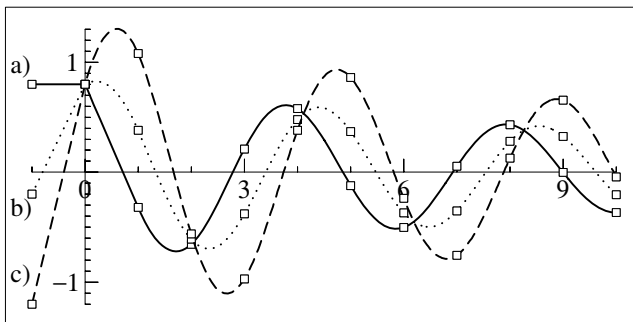


Fig. 17.2. Solutions of (17.4)

Constant Step Size Methods for Constant Delay

If we apply the Runge-Kutta method (1.8) (or (7.7)) to a delay equation (17.1) we obtain

$$g_i^{(n)} = y_n + h \sum_j a_{ij} f(x_n + c_j h, g_j^{(n)}, y(x_n + c_j h - \tau))$$

$$y_{n+1} = y_n + h \sum_j b_j f(x_n + c_j h, g_j^{(n)}, y(x_n + c_j h - \tau)).$$

But which values should we give to $y(x_n + c_j h - \tau)$? If the delay is constant and satisfies $\tau = kh$ for some integer k , the most natural idea is to use the back-values of the old solution

$$g_i^{(n)} = y_n + h \sum_j a_{ij} f(x_n + c_j h, g_j^{(n)}, \gamma_j^{(n)}) \quad (17.5a)$$

$$y_{n+1} = y_n + h \sum_j b_j f(x_n + c_j h, g_j^{(n)}, \gamma_j^{(n)}) \quad (17.5b)$$

where

$$\gamma_j^{(n)} = \begin{cases} \varphi(x_n + c_j h - \tau) & \text{if } n < k \\ g_j^{(n-k)} & \text{if } n \geq k. \end{cases} \quad (17.5c)$$

This can be interpreted as solving successively

$$y'(x) = f(x, y(x), \varphi(x - \tau)) \quad (17.1a)$$

for the interval $[x_0, x_0 + \tau]$, then

$$\begin{aligned} y'(x) &= f(x, y(x), z(x)) \\ z'(x) &= f(x - \tau, z(x), \varphi(x - 2\tau)) \end{aligned} \quad (17.1b)$$

for the interval $[x_0 + \tau, x_0 + 2\tau]$, then

$$\begin{aligned} y'(x) &= f(x, y(x), z(x)) \\ z'(x) &= f(x - \tau, z(x), v(x)) \\ v'(x) &= f(x - 2\tau, v(x), \varphi(x - 3\tau)) \end{aligned} \quad (17.1c)$$

for the interval $[x_0 + 2\tau, x_0 + 3\tau]$, and so on. This is the perfect numerical analog of the “method of steps” mentioned above.

Theorem 17.1. *If c_i, a_{ij}, b_j are the coefficients of a p -th order Runge-Kutta method, then (17.5) is convergent of order p .*

Proof. The sequence (17.1a), (17.1b), ... are ordinary differential equations normally solved by a p th order Runge-Kutta method. Therefore the result follows immediately from Theorem 3.6. \square

Remark. For the collocation method based on Gaussian quadrature formula, Theorem 17.1 yields superconvergence in spite of the use of the low order approximations $\gamma_j^{(n)}$ of (17.5c). Bellen (1984) generalizes this result to the situation where $\tau = \tau(x)$ and $\gamma_j^{(n)}$ is the value of the collocation polynomial at $x_n + c_j h - \tau(x_n + c_j h)$. He proves superconvergence if the grid-points are chosen such that every interval $[x_{n-1}, x_n]$ is mapped, by $x - \tau(x)$, into $[x_{j-1}, x_j]$ for some $j < n$.

Numerical Example. We have integrated the problem

$$y'(x) = (1.4 - y(x-1)) \cdot y(x)$$

(see (17.12) below) for $0 \leq x \leq 10$ with initial values $y(x) = 0$, $-1 \leq x < 0$, $y(0) = 0.1$, and step sizes $h = 1, 1/2, 1/4, 1/8, \dots, 1/128$ using Kutta's methods of order 4 (Table 1.2, left). The absolute value of the global errors (and the solution in grey) are presented in Fig. 17.3. The 4th order convergence can clearly be observed. The downward peaks are provoked by sign changes in the error.

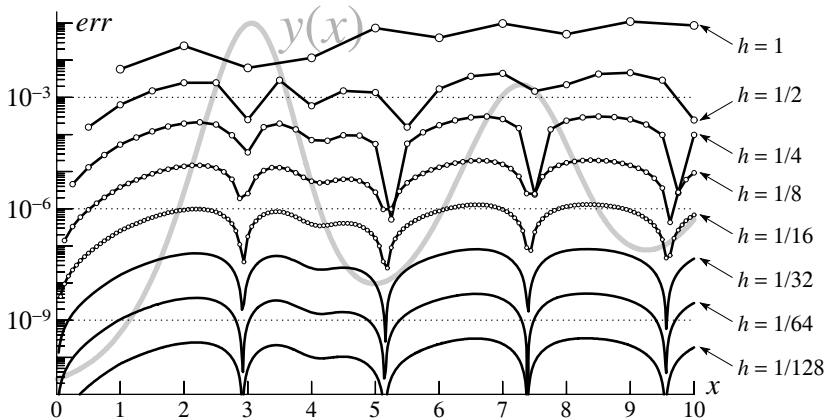


Fig. 17.3. Errors of RK44 with retarded stages (17.5)

Variable Step Size Methods

Although method (17.5) allows efficient and easy to code computations for simple problems with constant delays (such as all the examples of this section), it does not allow to change the step size arbitrarily, and an application to variable delay equations is not straightforward. If complete flexibility is desired, we need a *global* approximation to the solution. Such global approximations are furnished by multi-step methods of Adams or BDF type (see Chapter III.1) or the modern Runge-Kutta methods which are constructed together with a dense output. The code RETARD of the appendix is a modification of the code DOPRI5 (method of Dormand &

Prince in Table 5.2 with Shampine's dense output; see (6.12), (6.13) and the subsequent discussion) in such a way that after every successful step of integration the coefficients of the continuous solution are written into memory. Back-values of the solution are then available by calling the function YLAG(I,X,PHI). For example, for problem (17.4) the subroutine FCN would read as

$$F(1) = -1.4D0 * YLAG(1, X - 1.D0, PHI).$$

As we have seen, the solutions possess discontinuities in the derivatives at several points, e.g. for (17.1) at $x_0 + \tau$, $x_0 + 2\tau$, $x_0 + 3\tau, \dots$ etc. Therefore the code RETARD provides a possibility to match given points of discontinuities exactly (specify IWORK(6) and WORK(11), ...) which improves precision and computation time.

Earlier Runge-Kutta codes for delay equations have been written by Oppelstrup (1976), Oberle & Pesch (1981) and Bellen & Zennaro (1985). Bock & Schlöder (1981) exploited the natural dense output of multistep methods.

Stability

It can be observed from Fig. 17.1 and Fig. 17.2 that the solutions, after the initial phase, seem to tend to something like $e^{\alpha x} \cos \beta(x - \delta)$. We now try to determine α and β . We study the equation

$$y'(x) = \lambda y(x) + \mu y(x - 1). \quad (17.6)$$

There is no loss of generality in supposing the delay $\tau = 1$, since any delay $\tau \neq 1$ can be reduced to $\tau = 1$ by a coordinate change.

We search for a solution of the form

$$y(x) = e^{\gamma x} \quad \text{where } \gamma = \alpha + i\beta. \quad (17.7)$$

Introducing this into (17.6) we obtain the following "characteristic equation" for γ

$$\gamma - \lambda - \mu e^{-\gamma} = 0, \quad (17.8)$$

which, for $\mu \neq 0$, possesses an infinity of solutions: in fact, if $|\gamma|$ becomes large, we obtain from (17.8), since λ is fixed, that $\mu e^{-\gamma}$ must be large too and

$$\gamma \approx \mu e^{-\gamma}. \quad (17.8')$$

This implies that $\gamma = \alpha + i\beta$ is close to the imaginary axis. Hence $|\gamma| \approx |\beta|$ and from (17.8')

$$|\beta| \approx |\mu| e^{-\alpha}.$$

Therefore the roots of (17.8) lie asymptotically on the curves $-\alpha = \log |\beta| - \log |\mu|$. Again from (17.8'), we have a root whenever the argument of $\mu e^{-i\beta}$ is close to $\pi/2$ (for $\beta > 0$), i.e. if

$$\beta \approx \arg \mu - \frac{\pi}{2} + 2k\pi \quad k = 1, 2, \dots$$

There are thus two sequences of characteristic values which tend to infinity on logarithmic curves left of the imaginary axis, with 2π as asymptotic distance between two consecutive values.

The “general solution” of (17.6) is thus a Fourier-like superposition of solutions of type (17.7) (Wright 1946, see also Bellman & Cooke 1963, Chapter 4). The larger $-\operatorname{Re} \gamma$ is, the faster these solutions “die out” as $x \rightarrow \infty$. The dominant solutions are thus (provided that the corresponding coefficients are not zero) those which correspond to the largest real part, i.e., those closest to the origin. For equations (17.3) and (17.4) the characteristic equations are $\gamma + e^{-\gamma} = 0$ and $\gamma + 1.4e^{-\gamma} = 0$ with solutions $\gamma = -0.31813 \pm 1.33724i$ and $\gamma = -0.08170 \pm 1.51699i$ respectively, which explains nicely the behaviour of the asymptotic solutions of Fig. 17.1 and Fig. 17.2.

Remark. For the case of *matrix equations*

$$y'(x) = Ay(x) + By(x-1)$$

where A and B are not simultaneously diagonalizable, we set $y(x) = ve^{\gamma x}$ where $v \neq 0$ is a given vector. The equation now leads to

$$\gamma v = Av + Be^{-\gamma}v,$$

which has a nontrivial solution if

$$\det(\gamma I - A - Be^{-\gamma}) = 0, \quad (17.8'')$$

the characteristic equation for the more general case. The shape of the solutions of (17.8'') is similar to those of (17.8), there are just $r = \operatorname{rank}(B)$ points in each strip of width 2π instead of one.

All solutions of (17.6) remain *stable* for $x \rightarrow \infty$ if all characteristic roots of (17.8) remain in the negative half plane. This result follows either from the above expansion theorem or from the theory of Laplace transforms (e.g., Bellmann & Cooke (1963), Chapter 1), which, in fact, is closely related.

In order to study the boundary of the stability domain, we search for (λ, μ) values for which the first solution γ crosses the imaginary axis, i.e. $\gamma = i\theta$ for θ real. If we insert this into (17.8), we obtain

$$\begin{aligned} \lambda &= -\mu & \text{for } \theta = 0 \ (\gamma \text{ real}) \\ \lambda &= i\theta - \mu e^{-i\theta} & \text{for } \theta \neq 0 \end{aligned}$$

or, by separating real and imaginary parts,

$$\lambda = \frac{\cos \theta \cdot \theta}{\sin \theta}, \quad \mu = -\frac{\theta}{\sin \theta}$$

valid for real λ and μ . These paths are sketched in Fig. 17.4 and separate in the (λ, μ) -plane the domains of stability and instability for the solutions of (17.6) (a result of Hayes 1950).

If we put $\theta = \pi/2$, we find that the solutions of $y'(x) = \mu y(x-1)$ remain *stable* for

$$-\frac{\pi}{2} \leq \mu \leq 0 \quad (17.9a)$$

and are *unstable* for

$$\mu < -\frac{\pi}{2} \quad \text{as well as} \quad \mu > 0. \quad (17.9b)$$

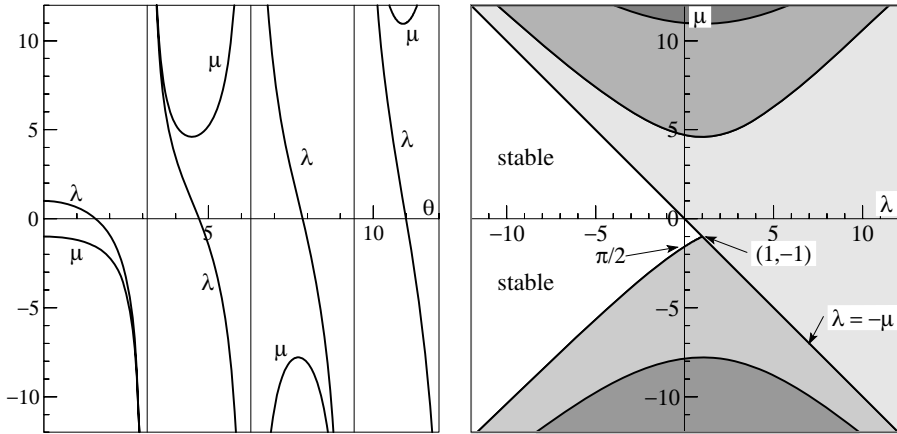


Fig. 17.4. Domain of stability for $y'(x) = \lambda y(x) + \mu y(x-1)$

An Example from Population Dynamics

Lord Cherwell drew my attention to an equation, equivalent to (8) (here: (17.12)) with $a = \log 2$, which he had encountered in his application of probability methods to the problem of distribution of primes. My thanks are due to him for thus introducing me to an interesting problem. (E.M. Wright 1945)

We now demonstrate the phenomena discussed above and the power of our programs on a couple of examples drawn from applications. For supplementary applications of delay equations to all sorts of sciences, consult the impressive list in Driver (1977, p. 239-240).

Let $y(x)$ represent the population of a certain species, whose development as a function of time is to be studied. The simple model of infinite exponential growth $y' = \lambda y$ was soon replaced by the hypothesis that the growth rate λ will decrease with increasing population y due to illness and lack of food and space. One then arrives at the model (Verhulst 1845, Pearl & Reed 1922)

$$y'(x) = k \cdot (a - y(x)) \cdot y(x). \quad (17.10)$$

“Nous donnerons le nom *logistique* à la courbe caractérisée par l’équation précédente” (Verhulst). It can be solved by elementary functions (Exercise 1). All solutions with initial value $y_0 > 0$ tend asymptotically to a as $x \rightarrow \infty$. If we assume the growth rate to depend on the population of the *preceding* generation, (17.10) becomes a delay equation (Cunningham 1954, Wright 1955, Kakutani & Markus 1958)

$$y'(x) = k \cdot (a - y(x - \tau)) \cdot y(x). \quad (17.11)$$

Introducing the new function $z(x) = k\tau y(\tau x)$ into (17.11) and again replacing z by y and $ka\tau$ by a we obtain

$$y'(x) = (a - y(x - 1)) \cdot y(x). \quad (17.12)$$

This equation has an equilibrium point at $y(x) = a$. The substitution $y(x) = a + z(x)$ and linearization leads to the equation $z'(x) = -az(x - 1)$, and condition (17.9) shows that this equilibrium point is locally stable if $0 < a \leq \pi/2$. Hence the characteristic equation, here $\gamma + ae^{-\gamma} = 0$, possesses two real solutions iff $a < 1/e = 0.368$, which makes monotonic solutions possible; otherwise they are oscillatory. For $a > \pi/2$ the equilibrium solution is unstable and gives rise to a periodic limit cycle.

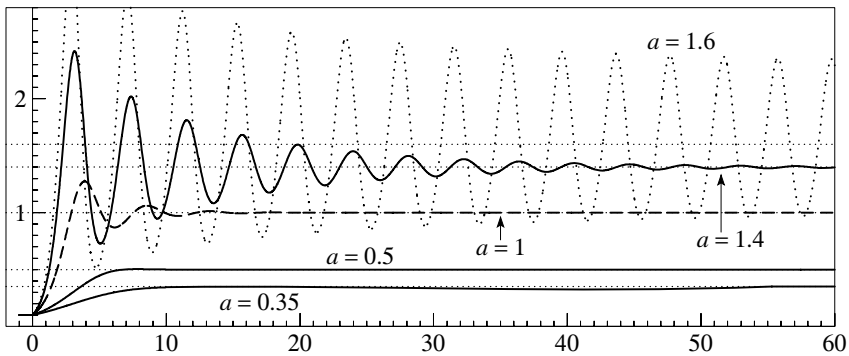


Fig. 17.5. Solutions of the population dynamics problem (17.12)

The solutions in Fig. 17.5 have been computed by the code RETARD of the appendix with subroutine FCN as

$$F(1) = (A - \text{YLAG}(1, X - 1.D0, \text{PHI})) * Y(1), \quad A = 0.35, 0.5, 1., 1.4, \text{ and } 1.6.$$

Infectious Disease Modelling

De tous ceux qui ont traité cette matière, c'est sans contredit M. de la Condamine qui l'a fait avec plus de succès. Il est déjà venu à bout de persuader la meilleure partie du monde raisonnable de la grande utilité de l'inoculation: quant aux autres, il serait inutile de vouloir employer la raison avec eux: puisqu'ils n'agissent pas par principes. Il faut les conduire comme des enfants vers leur mieux ...
(Daniel Bernoulli 1760)

Daniel Bernoulli ("Docteur en medecine, Professeur de Physique en l'Université de Bâle, Associé étranger de l'Academie des Sciences") was the first to use differential calculus to model infectious diseases in his 1760 paper on smallpox vaccination. At the beginning of our century, mathematical modelling of epidemics gained new interest. This finally led to the classical model of Kermack & McKendrick (1927): let $y_1(x)$ measure the *susceptible* portion of the population, $y_2(x)$ the *infected*, and $y_3(x)$ the *removed* (e.g. immunized) one. It is then natural to assume that the number of newly infected people per time unit is proportional to the product $y_1(x)y_2(x)$, just as in bimolecular chemical reactions (see Section I.16). If we finally assume the number of newly removed persons to be proportional to the infected ones, we arrive at the model

$$y_1' = -y_1 y_2, \quad y_2' = y_1 y_2 - y_2, \quad y_3' = y_2 \quad (17.13)$$

where we have taken for simplicity all rate constants equal to one. This system can be integrated by elementary methods (divide the first two equations and solve $dy_2/dy_1 = -1 + 1/y_1$). The numerical solution with initial values $y_1(0) = 5$, $y_2(0) = 0.1$, $y_3(0) = 1$ is painted in gray color in Fig. 17.6: an epidemic breaks out, everybody finally becomes "removed" and nothing further happens.

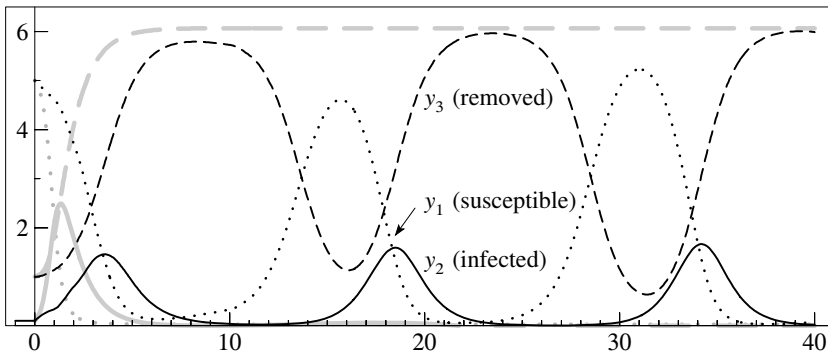


Fig. 17.6. Periodic outbreak of disease, model (17.14)
(in gray: Solution of Kermack - McKendrick model (17.13))

We arrive at a periodic outbreak of the disease, if we assume that immunized people become susceptible again, say after a fixed time τ ($\tau = 10$). If we also

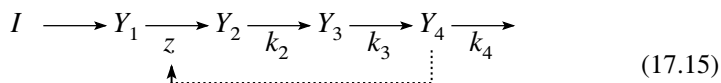
introduce an incubation period of, say, $\tau_2 = 1$, we arrive at the model

$$\begin{aligned}y_1'(x) &= -y_1(x)y_2(x-1) + y_2(x-10) \\y_2'(x) &= y_1(x)y_2(x-1) - y_2(x) \\y_3'(x) &= y_2(x) - y_2(x-10)\end{aligned}\tag{17.14}$$

instead of (17.13). The solutions of (17.14), for the initial phases $y_1(x) = 5$, $y_2(x) = 0.1$, $y_3(x) = 1$ for $x \leq 0$, are shown in Fig. 17.6 and illustrate the periodic outbreak of the disease.

An Example from Enzyme Kinetics

Our next example, more complicated than the preceding ones, is from enzyme kinetics (Okamoto & Hayashi 1984). Consider the following consecutive reactions



where I is an exogenous substrate supply which is maintained constant and n molecules of the end product Y_4 inhibit co-operatively the reaction step of $Y_1 \rightarrow Y_2$ as

$$z = \frac{k_1}{1 + \alpha(y_4(x))^n}.$$

It is generally expected that the inhibitor molecule must be moved to the position of the regulatory enzyme by forces such as diffusion or active transport. Thus, we consider this time consuming process causing time-delay and we arrive at the model

$$\begin{aligned}y_1'(x) &= I - zy_1(x) \\y_2'(x) &= zy_1(x) - y_2(x) \\y_3'(x) &= y_2(x) - y_3(x) \\y_4'(x) &= y_3(x) - 0.5y_4(x)\end{aligned}\quad z = \frac{1}{1 + 0.0005(y_4(x-4))^3}.\tag{17.16}$$

This system possesses an equilibrium at $zy_1 = y_2 = y_3 = I$, $y_4 = 2I$, $y_1 = I(1 + 0.004I^3) =: c_1$. When it is linearized in the neighbourhood of this equilibrium point, it becomes

$$\begin{aligned}y_1'(x) &= -c_1y_1(x) + c_2y_4(x-4) \\y_2'(x) &= c_1y_1(x) - y_2(x) - c_2y_4(x-4) \\y_3'(x) &= y_2(x) - y_3(x) \\y_4'(x) &= y_3(x) - 0.5y_4(x)\end{aligned}\tag{17.17}$$

where $c_2 = c_1 \cdot I^3 \cdot 0.006$. By setting $y(x) = v \cdot e^{\gamma x}$ we arrive at the characteristic equation (see (17.8')), which becomes after some simplifications

$$(c_1 + \gamma)(1 + \gamma)^2(0.5 + \gamma) + c_2\gamma e^{-4\gamma} = 0. \quad (17.18)$$

As in the paper of Okamoto & Hayashi, we put $I = 10.5$. Then (17.18) possesses one pair of complex solutions in \mathbb{C}^+ , namely

$$\gamma = 0.04246 \pm 0.47666i$$

and the equilibrium solution is unstable (see Fig. 17.7). The period of the solution of the linearized equation is thus $T = 2\pi/0.47666 = 13.18$. The solutions then tend to a limit cycle of approximately the same period.

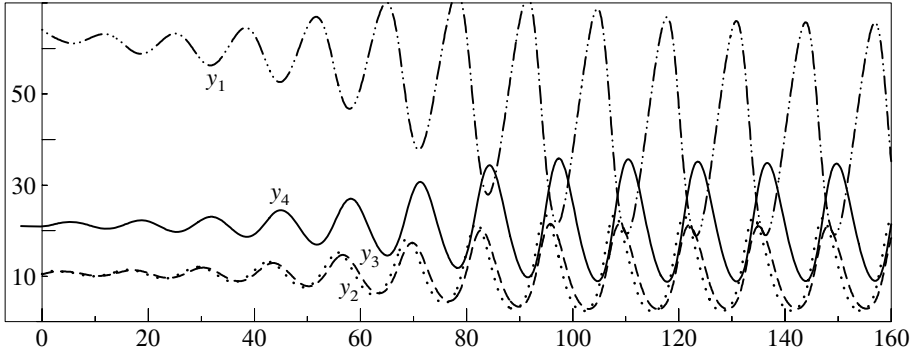


Fig. 17.7. Solutions of the enzyme kinetics problem (17.16), $I = 10.5$.
Initial values close to equilibrium position

A Mathematical Model in Immunology

We conclude our series of examples with Marchuk's model (Marchuk 1975) for the struggle of viruses $V(t)$, antibodies $F(t)$ and plasma cells $C(t)$ in the organism of a person infected by a viral disease. The equations are

$$\begin{aligned} \frac{dV}{dt} &= (h_1 - h_2 F)V \\ \frac{dC}{dt} &= \xi(m)h_3 F(t - \tau)V(t - \tau) - h_5(C - 1) \\ \frac{dF}{dt} &= h_4(C - F) - h_8 FV. \end{aligned} \quad (17.19)$$

The first is a Volterra - Lotka like predator-prey equation. The second equation describes the creation of new plasma cells with time lag due to infection, in the absence of which the second term creates an equilibrium at $C = 1$. The third equation models the creation of antibodies from plasma cells ($h_4 C$) and their

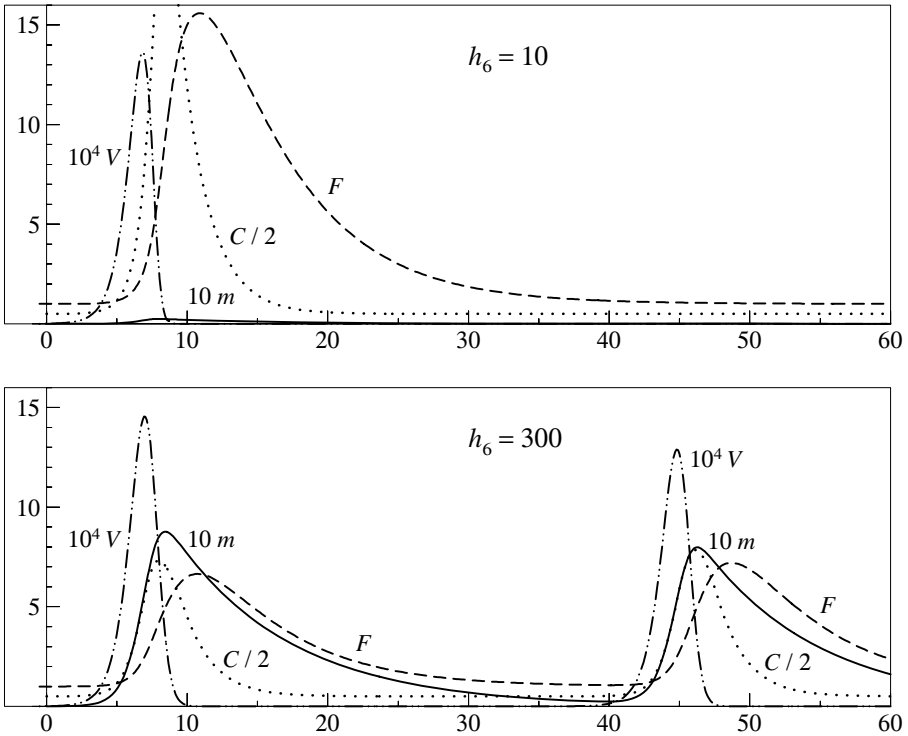


Fig. 17.8. Solutions of the Marchuk immunology model

decrease due to aging ($-h_4 F$) and binding with antigens ($-h_8 FV$). The term $\xi(m)$, finally, is defined by

$$\xi(m) = \begin{cases} 1 & \text{if } m \leq 0.1 \\ (1-m)\frac{10}{9} & \text{if } 0.1 \leq m \leq 1 \end{cases}$$

and expresses the fact that the creation of plasma cells slows down when the organism is damaged by the viral infection. The relative characteristic $m(t)$ of damaging is given by a fourth equation

$$\frac{dm}{dt} = h_6 V - h_7 m$$

where the first term expresses the damaging and the second recuperation.

This model allows us, by changing the coefficients h_1, h_2, \dots, h_8 , to model all sorts of behaviour of stable health, unstable health, acute form of a disease, chronic form etc. See Chapter 2 of Marchuk (1983). In Fig. 17.8 we plot the solutions of this model for $\tau = 0.5$, $h_1 = 2$, $h_2 = 0.8$, $h_3 = 10^4$, $h_4 = 0.17$, $h_5 = 0.5$, $h_7 = 0.12$, $h_8 = 8$ and initial values $V(t) = \max(0, 10^{-6} + t)$ if $t \leq 0$, $C(0) = 1$, $F(t) = 1$ if $t \leq 0$, $m(0) = 0$. In dependence of the value of h_6 ($h_6 = 10$

or $h_6 = 300$), we then observe either complete recovery (defined by $V(t) < 10^{-16}$), or periodic outbreak of the disease due to damaging ($m(t)$ becomes nearly 1).

Integro-Differential Equations

Often the hypothesis that a system depends on the time lagged solution at a specified fixed value $x - \tau$ is not very realistic, and one should rather suppose this dependence to be stretched out over a longer period of time. Then, instead of (17.1), we would have for example

$$y'(x) = f\left(x, y(x), \int_{x-\tau}^x K(x, \xi, y(\xi)) d\xi\right). \quad (17.20)$$

The numerical treatment of these problems becomes much more expensive (see Brunner & van der Houwen (1986) for a study of various discretization methods). If $K(x, \xi, y)$ is zero in the neighbourhood of the diagonal $x = \xi$, one can eventually use RETARD and call a quadrature routine for each function evaluation.

Fortunately, many integro-differential equations can be reduced to ordinary or delay differential equations by introducing new variables for the integral function.

Example (Volterra 1934). Consider the equation

$$y'(x) = \left(\varepsilon - \alpha y(x) - \int_0^x k(x - \xi)y(\xi) d\xi\right) \cdot y(x) \quad (17.21)$$

for population dynamics, where the integral term represents a decrease of the reproduction rate due to pollution. If now for example $k(x) = c$, we put

$$\int_0^x y(\xi) d\xi = v(x), \quad y(x) = v'(x)$$

and obtain

$$v''(x) = (\varepsilon - \alpha v'(x) - cv(x)) \cdot v'(x),$$

an ordinary differential equation.

The same method is possible for equations (17.20) with “degenerate kernel”; i.e., where

$$K(x, \xi, y) = \sum_{i=1}^m a_i(x)b_i(\xi, y). \quad (17.22)$$

If we insert this into (17.20) and put

$$v_i(x) = \int_{x-\tau}^x b_i(\xi, y(\xi)) d\xi, \quad (17.23)$$

we obtain

$$\begin{aligned} y'(x) &= f\left(x, y(x), \sum_{i=1}^m a_i(x) v_i(x)\right) \\ v'_i(x) &= b_i(x, y(x)) - b_i(x - \tau, y(x - \tau)) \quad i = 1, \dots, m, \end{aligned} \quad (17.20')$$

a system of delay differential equations.

Exercises

1. Compute the solution of the Verhulst & Pearl equation (17.10).
2. Compute the equilibrium points of Marchuk's equation (17.19) and study their stability.
3. Assume that the kernel $k(x)$ in Volterra's equation (17.21) is given by

$$k(x) = p(x)e^{-\beta x}$$

where $p(x)$ is some polynomial. Show that this problem can be transformed into an ordinary differential equation.

4. Consider the integro-differential equation

$$y'(x) = f\left(x, y(x), \int_0^x K(x, \xi, y(\xi)) d\xi\right). \quad (17.24)$$

- a) For the degenerate kernel (17.22) problem (17.24) becomes equivalent to the ordinary differential equation

$$\begin{aligned} y'(x) &= f\left(x, y(x), \sum_{j=1}^m a_j(x) v_j(x)\right) \\ v'_j(x) &= b_j(x, y(x)). \end{aligned} \quad (17.25)$$

- b) Show that an application of an explicit (p th order) Runge-Kutta method to (17.25) yields the formulas (Pouzet 1963)

$$\begin{aligned} y_{n+1} &= y_n + h \sum_{i=1}^s b_i f(x_n + c_i h, g_i^{(n)}, u_i^{(n)}) \\ g_i^{(n)} &= y_n + h \sum_{j=1}^{i-1} a_{ij} f(x_n + c_j h, g_j^{(n)}, u_j^{(n)}) \\ u_i^{(n)} &= F_n(x_n + c_i h) + h \sum_{j=1}^{i-1} a_{ij} K(x_n + c_i h, x_n + c_j h, g_j^{(n)}) \end{aligned} \quad (17.26)$$

where

$$F_0(x) = 0, \quad F_{n+1}(x) = F_n(x) + h \sum_{i=1}^s b_i K(x, x_n + c_i h, g_i^{(n)}).$$

- c) If we apply method (17.26) to problem (17.24), where the kernel does not necessarily satisfy (17.22), we nevertheless have convergence of order p .

Hint. Approximate the kernel by a degenerate one.

5. (Zennaro 1986). For the delay equation (17.1) consider the method (17.5) where (17.5c) is replaced by

$$\gamma_j^{(n)} = \begin{cases} \varphi(x_n + c_j h - \tau) & \text{if } n < k \\ q_{n-k}(c_j) & \text{if } n \geq k. \end{cases} \quad (17.5c')$$

Here $q_n(\theta)$ is the polynomial given by a continuous Runge-Kutta method (Section II.6)

$$q_n(\theta) = y_n + h \sum_{j=1}^s b_j(\theta) f(x_n + c_j h, g_j^{(n)}, \gamma_j^{(n)}).$$

- a) Prove that the orthogonality conditions

$$\int_0^1 \theta^{q-1} \left(\gamma(t) \sum_{j=1}^s b_j(\theta) \Phi_j(t) - \theta^{\varrho(t)} \right) d\theta = 0 \quad \text{for } q + \varrho(t) \leq p \quad (17.27)$$

imply convergence of order p , if the underlying Runge-Kutta method is of order p for ordinary differential equations.

Hint. Use the theory of B-series and the Gröbner - Alekseev formula (I.14.18) of Section I.14.

- b) If for a given Runge-Kutta method the polynomials $b_j(\theta)$ of degree $\leq [(p+1)/2]$ are such that $b_j(0) = 0$, $b_j(1) = b_j$ and

$$\int_0^1 \theta^{q-1} b_j(\theta) d\theta = \frac{1}{q} b_j(1 - c_j^q), \quad q = 1, \dots, [(p-1)/2], \quad (17.28)$$

then (17.27) is satisfied. In addition one has the order conditions

$$\sum_{j=1}^s b_j(\theta) \Phi_j(t) = \frac{\theta^{\varrho(t)}}{\gamma(t)} \quad \text{for } \varrho(t) \leq [(p+1)/2].$$

- c) Show that the conditions (17.28) admit unique polynomials $b_j(\theta)$ of degree $[(p+1)/2]$.
6. Solve Volterra's equation (17.21) with $k(x) = c$ and compare the solution with the "pollution free" problem (17.10). Which population lives better, that *with* pollution, or that without?

Chapter III. Multistep Methods and General Linear Methods

This chapter is devoted to the study of multistep and general multivalued methods. After retracing their historical development (Adams, Nyström, Milne, BDF) we study in the subsequent sections the order, stability and convergence properties of these methods. Convergence is most elegantly set in the framework of one-step methods in higher dimensions. Sections III.5 and III.6 are devoted to variable step size and Nordsieck methods. We then discuss the various available codes and compare them on the numerical examples of Section II.10 as well as on some equations of high dimension. Before closing the chapter with a section on special methods for second order equations, we discuss two highly theoretical subjects: one on general linear methods, including Runge-Kutta methods as well as multistep methods and many generalizations, and the other on the asymptotic expansion of the global error of such methods.

III.1 Classical Linear Multistep Formulas

... , and my undertaking must have ended here, if I had depended upon my own resources. But at this point Professor J.C. Adams furnished me with a perfectly satisfactory method of calculating by quadratures the exact theoretical forms of drops of fluids from the Differential Equation of Laplace, ... (F. Bashforth 1883)

Another improvement of Euler's method was considered even earlier than Runge-Kutta methods — the methods of Adams. These were devised by John Couch Adams in order to solve a problem of F. Bashforth, which occurred in an investigation of capillary action. Both the problem and the numerical integration schemes are published in Bashforth (1883). The actual origin of these methods must date back to at least 1855, since in that year F. Bashforth made an application to the Royal Society for assistance from the Government grant. There he wrote: "... , but I am indebted to Mr Adams for a method of treating the differential equation

$$\frac{\frac{ddz}{du^2}}{\left(1 + \frac{dz^2}{du^2}\right)^{3/2}} + \frac{\frac{1}{u} \frac{dz}{du}}{\left(1 + \frac{dz^2}{du^2}\right)^{1/2}} - 2\alpha z = \frac{2}{b},$$

when put under the form

$$\frac{b}{\varrho} + \frac{b}{x} \sin \varphi = 2 + 2\alpha b^2 \frac{z}{b} = 2 + \beta \frac{z}{b},$$

which gives the theoretical form of the drop with an accuracy exceeding that of the most refined measurements."

In contrast to one-step methods, where the numerical solution is obtained solely from the differential equation and the initial value, the algorithm of Adams consists of two parts: firstly, a *starting procedure* which provides y_1, \dots, y_{k-1} (approximations to the exact solution at the points $x_0 + h, \dots, x_0 + (k-1)h$) and, secondly, a *multistep formula* to obtain an approximation to the exact solution $y(x_0 + kh)$. This is then applied recursively, based on the numerical approximations of k successive steps, to compute $y(x_0 + (k+1)h)$, etc.

There are several possibilities for obtaining the missing starting values. J.C. Adams actually computed them using the Taylor series expansion of the exact solution (as described in Section I.8, see also Exercise 2). Another possibility is the use of any one-step method, e.g., a Runge-Kutta method (see Chapter II). It is also usual to start with low-order Adams methods and very small step sizes.

Explicit Adams Methods

We now derive, following Adams, the first explicit multistep formulas. We introduce the notation $x_i = x_0 + ih$ for the grid points and suppose we know the numerical approximations $y_n, y_{n-1}, \dots, y_{n-k+1}$ to the exact solution $y(x_n), \dots, y(x_{n-k+1})$ of the differential equation

$$y' = f(x, y), \quad y(x_0) = y_0. \quad (1.1)$$

Adams considers (1.1) in integrated form,

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(t, y(t)) dt. \quad (1.2)$$

On the right hand side of (1.2) there appears the unknown solution $y(x)$. But since the approximations y_{n-k+1}, \dots, y_n are known, the values

$$f_i = f(x_i, y_i) \quad \text{for } i = n-k+1, \dots, n \quad (1.3)$$

are also available and it is natural to replace the function $f(t, y(t))$ in (1.2) by the interpolation polynomial through the points $\{(x_i, f_i) \mid i = n-k+1, \dots, n\}$ (see Fig. 1.1).

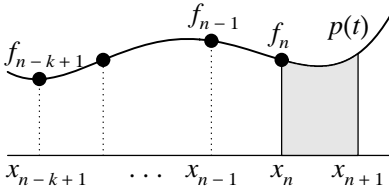


Fig. 1.1. Explicit Adams methods

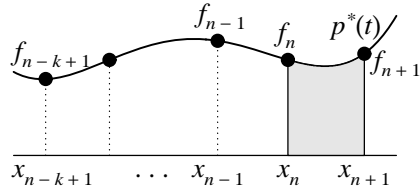


Fig. 1.2. Implicit Adams methods

This polynomial can be expressed in terms of backward differences

$$\nabla^0 f_n = f_n, \quad \nabla^{j+1} f_n = \nabla^j f_n - \nabla^j f_{n-1}$$

as follows:

$$p(t) = p(x_n + sh) = \sum_{j=0}^{k-1} (-1)^j \binom{-s}{j} \nabla^j f_n \quad (1.4)$$

(Newton's interpolation formula of 1676, published in Newton (1711), see e.g. Henrici (1962), p. 190). The numerical analogue to (1.2) is then given by

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} p(t) dt$$

or after insertion of (1.4) by

$$y_{n+1} = y_n + h \sum_{j=0}^{k-1} \gamma_j \nabla^j f_n \quad (1.5)$$

where the coefficients γ_j satisfy

$$\gamma_j = (-1)^j \int_0^1 \binom{-s}{j} ds \quad (1.6)$$

(see Table 1.1 for their numerical values). A simple recurrence relation for these coefficients will be derived below (formula (1.7)).

Table 1.1. Coefficients for the explicit Adams methods

j	0	1	2	3	4	5	6	7	8
γ_j	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$	$\frac{95}{288}$	$\frac{19087}{60480}$	$\frac{5257}{17280}$	$\frac{1070017}{3628800}$

Special cases of (1.5). For $k = 1, 2, 3, 4$, after expressing the backward differences in terms of f_{n-j} , one obtains the formulas

$$\begin{aligned} k=1: \quad & y_{n+1} = y_n + hf_n && \text{(explicit Euler method)} \\ k=2: \quad & y_{n+1} = y_n + h \left(\frac{3}{2}f_n - \frac{1}{2}f_{n-1} \right) \\ k=3: \quad & y_{n+1} = y_n + h \left(\frac{23}{12}f_n - \frac{16}{12}f_{n-1} + \frac{5}{12}f_{n-2} \right) \\ k=4: \quad & y_{n+1} = y_n + h \left(\frac{55}{24}f_n - \frac{59}{24}f_{n-1} + \frac{37}{24}f_{n-2} - \frac{9}{24}f_{n-3} \right). \end{aligned} \quad (1.5')$$

Recurrence relation for the coefficients. Using Euler's method of *generating functions* we can deduce a simple recurrence relation for γ_i (see e.g. Henrici 1962). Denote by $G(t)$ the series

$$G(t) = \sum_{j=0}^{\infty} \gamma_j t^j.$$

With the definition of γ_j and the binomial theorem one obtains

$$\begin{aligned} G(t) &= \sum_{j=0}^{\infty} (-t)^j \int_0^1 \binom{-s}{j} ds = \int_0^1 \sum_{j=0}^{\infty} (-t)^j \binom{-s}{j} ds \\ &= \int_0^1 (1-t)^{-s} ds = -\frac{t}{(1-t) \log(1-t)}. \end{aligned}$$

This can be written as

$$-\frac{\log(1-t)}{t} G(t) = \frac{1}{1-t}$$

or as

$$\left(1 + \frac{1}{2}t + \frac{1}{3}t^2 + \dots\right) (\gamma_0 + \gamma_1 t + \gamma_2 t^2 + \dots) = (1 + t + t^2 + \dots).$$

Comparing the coefficients of t^m we get the desired recurrence relation

$$\gamma_m + \frac{1}{2}\gamma_{m-1} + \frac{1}{3}\gamma_{m-2} + \dots + \frac{1}{m+1}\gamma_0 = 1. \quad (1.7)$$

Implicit Adams Methods

The formulas (1.5) are obtained by integrating the interpolation polynomial (1.4) from x_n to x_{n+1} , i.e., outside the interpolation interval (x_{n-k+1}, x_n) . It is well known that an interpolation polynomial is usually a rather poor approximation outside this interval. Adams therefore also investigated methods where (1.4) is replaced by the interpolation polynomial which uses in addition the point (x_{n+1}, f_{n+1}) , i.e.,

$$p^*(t) = p^*(x_n + sh) = \sum_{j=0}^k (-1)^j \binom{-s+1}{j} \nabla^j f_{n+1} \quad (1.8)$$

(see Fig. 1.2). Inserting this into (1.2) we obtain the following implicit method

$$y_{n+1} = y_n + h \sum_{j=0}^k \gamma_j^* \nabla^j f_{n+1} \quad (1.9)$$

where the coefficients γ_j^* satisfy

$$\gamma_j^* = (-1)^j \int_0^1 \binom{-s+1}{j} ds \quad (1.10)$$

and are given in Table 1.2 for $j \leq 8$. Again, a simple recurrence relation can be derived for these coefficients (Exercise 3).

Table 1.2. Coefficients for the implicit Adams methods

j	0	1	2	3	4	5	6	7	8
γ_j^*	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$	$-\frac{863}{60480}$	$-\frac{275}{24192}$	$-\frac{33953}{3628800}$

The formulas thus obtained are generally of the form

$$y_{n+1} = y_n + h(\beta_k f_{n+1} + \dots + \beta_0 f_{n-k+1}). \quad (1.9')$$

The first examples are as follows

$$\begin{aligned}
 k = 0: \quad y_{n+1} &= y_n + hf_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) \\
 k = 1: \quad y_{n+1} &= y_n + h\left(\frac{1}{2}f_{n+1} + \frac{1}{2}f_n\right) \\
 k = 2: \quad y_{n+1} &= y_n + h\left(\frac{5}{12}f_{n+1} + \frac{8}{12}f_n - \frac{1}{12}f_{n-1}\right) \\
 k = 3: \quad y_{n+1} &= y_n + h\left(\frac{9}{24}f_{n+1} + \frac{19}{24}f_n - \frac{5}{24}f_{n-1} + \frac{1}{24}f_{n-2}\right).
 \end{aligned} \tag{1.9''}$$

The special cases $k = 0$ and $k = 1$ are the implicit Euler method and the trapezoidal rule, respectively. They are actually one-step methods and have already been considered in Chapter II.7.

The methods (1.9) give in general more accurate approximations to the exact solution than (1.5). This will be discussed in detail when the concepts of order and error constant are introduced (Section III.2). The price for this higher accuracy is that y_{n+1} is only defined implicitly by formula (1.9). Therefore, in general a nonlinear equation has to be solved at each step.

Predictor-corrector methods. One possibility for solving this nonlinear equation is to apply fixed point iteration. In practice one proceeds as follows:

- P: compute the predictor $\hat{y}_{n+1} = y_n + h \sum_{j=0}^{k-1} \gamma_j \nabla^j f_n$ by the explicit Adams method (1.5); this already yields a reasonable approximation to $y(x_{n+1})$;
- E: evaluate the function at this approximation: $\hat{f}_{n+1} = f(x_{n+1}, \hat{y}_{n+1})$;
- C: apply the corrector formula

$$y_{n+1} = y_n + h(\beta_k \hat{f}_{n+1} + \beta_{k-1} f_n + \dots + \beta_0 f_{n-k+1}) \tag{1.11}$$

to obtain y_{n+1} .

- E: evaluate the function anew, i.e., compute $f_{n+1} = f(x_{n+1}, y_{n+1})$.

This is the most common procedure, denoted by PECE. Other possibilities are: PECECE (two fixed point iterations per step) or PEC (one uses \hat{f}_{n+1} instead of f_{n+1} in the subsequent steps).

This predictor-corrector technique has been used by F.R. Moulton (1926) as well as by W.E. Milne (1926). J.C. Adams actually solved the implicit equation (1.9) by Newton's method, in the same way as is now usual for stiff equations (see Volume II).

Remark. Formula (1.5) is often attributed to Adams-Bashforth. Similarly, the multistep formula (1.9) is usually attributed to Adams-Moulton (Moulton 1926). In fact, both formulas are due to Adams.

Numerical Experiment

We consider the Van der Pol equation (I.16.2) with $\varepsilon = 1$, take as initial values $y_1(0) = A$, $y_2(0) = 0$ on the limit cycle and integrate over one period T (for the values of A and T see Exercise I.16.1). This is exactly the same problem as the one used for the comparison of Runge-Kutta methods (Fig. II.1.1). We have applied the above explicit and implicit Adams methods with several fixed step sizes. The missing starting values were computed with high accuracy by an explicit Runge-Kutta method. Fig. 1.3 shows the errors of both components in dependence of the number of function evaluations. Since we have implemented the implicit method (1.9) in PECE mode it requires 2 function evaluations per step, whereas the explicit method (1.5) needs only one.

This experiment shows that, for the same value of k , the implicit methods usually give a better result (the strange behaviour in the error of the y_2 -component for $k \geq 3$ is due to a sign change). Since we have used double logarithmic scales, it is possible to read the “numerical order” from the slope of the corresponding lines. We observe that the global error of the explicit Adams methods behaves like $\mathcal{O}(h^k)$ and that of the implicit methods like $\mathcal{O}(h^{k+1})$. This will be proved in the following sections.

We also remark that the scales used in Fig. 1.3 are exactly the same as those of Fig. II.1.1. This allows a comparison with the Runge-Kutta methods of Section II.1.

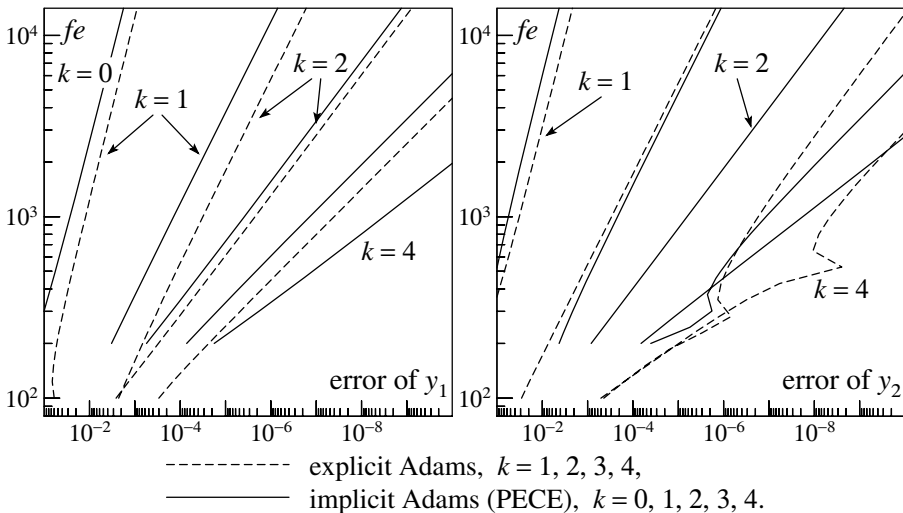


Fig. 1.3. Global errors versus number of function evaluations

Explicit Nyström Methods

Die angenäherte Integration hat, besonders in der letzten Zeit, ein ausgedehntes Anwendungsgebiet innerhalb der exakten Wissenschaften und der Technik gefunden. (E.J. Nyström 1925)

In his review article on the numerical integration of differential equations (which we have already encountered in Section II.14), Nyström (1925) also presents a new class of multistep methods. He considers instead of (1.2) the integral equation

$$y(x_{n+1}) = y(x_{n-1}) + \int_{x_{n-1}}^{x_{n+1}} f(t, y(t)) dt. \tag{1.12}$$

In the same way as above he replaces the unknown function $f(t, y(t))$ by the polynomial $p(t)$ of (1.4) and so obtains the formula (see Fig. 1.4)

$$y_{n+1} = y_{n-1} + h \sum_{j=0}^{k-1} \kappa_j \nabla^j f_n \tag{1.13}$$

with the coefficients

$$\kappa_j = (-1)^j \int_{-1}^1 \binom{-s}{j} ds. \tag{1.14}$$

The first of these coefficients are given in Table 1.3. E.J. Nyström recommended the formulas (1.13), because the coefficients κ_j were more convenient for his computations than the coefficients γ_j of (1.6). This recommendation, surely reasonable for a computation by hand, is of little relevance for computations on a computer.

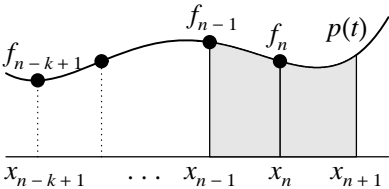


Fig. 1.4. Explicit Nyström methods

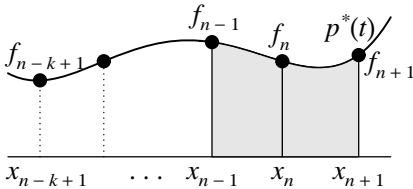


Fig. 1.5. Milne-Simpson methods

Table 1.3. Coefficients for the explicit Nyström methods

j	0	1	2	3	4	5	6	7	8
κ_j	2	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{29}{90}$	$\frac{14}{45}$	$\frac{1139}{3780}$	$\frac{41}{140}$	$\frac{32377}{113400}$

Special cases. For $k = 1$ the formula

$$y_{n+1} = y_{n-1} + 2hf_n \tag{1.13'}$$

is obtained. It is called the *mid-point rule* and is the simplest two-step method. Its symmetry was extremely useful in the extrapolation schemes of Section II.9. The case $k = 2$ yields nothing new, because $\kappa_1 = 0$. For $k = 3$ one gets

$$y_{n+1} = y_{n-1} + h \left(\frac{7}{3} f_n - \frac{2}{3} f_{n-1} + \frac{1}{3} f_{n-2} \right). \quad (1.13'')$$

Milne–Simpson Methods

We consider again the integral equation (1.12). But now we replace the integrand by the polynomial $p^*(t)$ of (1.8), which in addition to f_n, \dots, f_{n-k+1} also interpolates the value f_{n+1} (see Fig. 1.5). Proceeding as usual, we get the implicit formulas

$$y_{n+1} = y_{n-1} + h \sum_{j=0}^k \kappa_j^* \nabla^j f_{n+1}. \quad (1.15)$$

The coefficients κ_j^* are defined by

$$\kappa_j^* = (-1)^j \int_{-1}^1 \binom{-s+1}{j} ds, \quad (1.16)$$

and the first of these are given in Table 1.4.

Table 1.4. Coefficients for the Milne–Simpson methods

j	0	1	2	3	4	5	6	7	8
κ_j^*	2	-2	$\frac{1}{3}$	0	$-\frac{1}{90}$	$-\frac{1}{90}$	$-\frac{37}{3780}$	$-\frac{8}{945}$	$-\frac{119}{16200}$

If the backward differences in (1.15) are expressed in terms of f_{n-j} , one obtains the following methods for special values of k :

$$\begin{aligned} k=0: \quad y_{n+1} &= y_{n-1} + 2h f_{n+1}, \\ k=1: \quad y_{n+1} &= y_{n-1} + 2h f_n, \end{aligned} \quad (1.15')$$

$$k=2: \quad y_{n+1} = y_{n-1} + h \left(\frac{1}{3} f_{n+1} + \frac{4}{3} f_n + \frac{1}{3} f_{n-1} \right),$$

$$k=4: \quad y_{n+1} = y_{n-1} + h \left(\frac{29}{90} f_{n+1} + \frac{124}{90} f_n + \frac{24}{90} f_{n-1} + \frac{4}{90} f_{n-2} - \frac{1}{90} f_{n-3} \right).$$

The special case $k = 0$ is just Euler's implicit method applied with step size $2h$. For $k = 1$ one obtains the previously derived mid-point rule. The particular case

$k = 2$ is an interesting method, known as the *Milne method* (Milne 1926, 1970, p. 66). It is a direct generalization of Simpson's rule.

Many other similar methods have been investigated. They are all based on an integral equation of the form

$$y(x_{n+1}) = y(x_{n-\ell}) + \int_{x_{n-\ell}}^{x_{n+1}} f(t, y(t)) dt, \quad (1.17)$$

where $f(t, y(t))$ is replaced either by the interpolating polynomial $p(t)$ (formula (1.4)) or by $p^*(t)$ (formula (1.8)). E.g., for $\ell = 3$ one obtains

$$y_{n+1} = y_{n-3} + h \left(\frac{8}{3} f_n - \frac{4}{3} f_{n-1} + \frac{8}{3} f_{n-2} \right). \quad (1.18)$$

This particular method has been used by Milne (1926) as a “predictor” for his method: in order to solve the implicit equation (1.15'), Milne uses one or two fixed-point iterations with the numerical value of (1.18) as starting point.

Methods Based on Differentiation (BDF)

“My name is Gear.” — “pardon?”
 “Gear, dshii, ii, ay, are.” — “Mr. Jiea?”
 (In a hotel of Paris)

The multistep formulas considered until now are all based on numerical integration, i.e., the integral in (1.17) is approximated numerically using some quadrature formula. The underlying idea of the following multistep formulas is totally different as they are based on the numerical differentiation of a given function.

Assume that the approximations y_{n-k+1}, \dots, y_n to the exact solution of (1.1) are known. In order to derive a formula for y_{n+1} we consider the polynomial $q(x)$ which interpolates the values $\{(x_i, y_i) \mid i = n - k + 1, \dots, n + 1\}$. As in (1.8) this polynomial can be expressed in terms of backward differences, namely

$$q(x) = q(x_n + sh) = \sum_{j=0}^k (-1)^j \binom{-s+1}{j} \nabla^j y_{n+1}. \quad (1.19)$$

The unknown value y_{n+1} will now be determined in such a way that the polynomial $q(x)$ satisfies the differential equation at at least one grid-point, i.e.,

$$q'(x_{n+1-r}) = f(x_{n+1-r}, y_{n+1-r}). \quad (1.20)$$

For $r = 1$ we obtain *explicit* formulas. For $k = 1$ and $k = 2$, these are equivalent to the explicit Euler method and the mid-point rule, respectively. The case $k = 3$ yields

$$\frac{1}{3} y_{n+1} + \frac{1}{2} y_n - y_{n-1} + \frac{1}{6} y_{n-2} = h f_n. \quad (1.21)$$

This formula, however, as well as those for $k > 3$, is unstable (see Section III.3) and therefore useless.

Much more interesting are the formulas one obtains when (1.20) is taken for $r = 0$ (see Fig. 1.6).

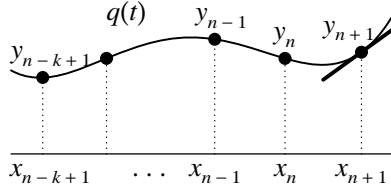


Fig. 1.6. Definition of BDF

In this case one gets the *implicit* formulas

$$\sum_{j=0}^k \delta_j^* \nabla^j y_{n+1} = h f_{n+1} \quad (1.22)$$

with the coefficients

$$\delta_j^* = (-1)^j \frac{d}{ds} \binom{-s+1}{j} \Big|_{s=1}.$$

Using the definition of the binomial coefficient

$$(-1)^j \binom{-s+1}{j} = \frac{1}{j!} (s-1)s(s+1) \dots (s+j-2)$$

the coefficients δ_j^* are obtained by direct differentiation:

$$\delta_0^* = 0, \quad \delta_j^* = \frac{1}{j} \quad \text{for } j \geq 1. \quad (1.23)$$

Formula (1.22) therefore becomes

$$\sum_{j=1}^k \frac{1}{j} \nabla^j y_{n+1} = h f_{n+1}. \quad (1.22')$$

These multistep formulas, known as *backward differentiation formulas* (or *BDF-methods*), are, since the work of Gear (1971), widely used for the integration of stiff differential equations (see Volume II). They were introduced by Curtiss & Hirschfelder (1952); Mitchell & Craggs (1953) call them “standard step-by-step methods”.

For the sake of completeness we give these formulas also in the form which expresses the backward differences in terms of the y_{n-j} .

$$\begin{aligned} k=1: \quad & y_{n+1} - y_n = h f_{n+1}, \\ k=2: \quad & \frac{3}{2} y_{n+1} - 2 y_n + \frac{1}{2} y_{n-1} = h f_{n+1}, \end{aligned} \quad (1.22'')$$

$$k = 3: \quad \frac{11}{6}y_{n+1} - 3y_n + \frac{3}{2}y_{n-1} - \frac{1}{3}y_{n-2} = hf_{n+1},$$

$$k = 4: \quad \frac{25}{12}y_{n+1} - 4y_n + 3y_{n-1} - \frac{4}{3}y_{n-2} + \frac{1}{4}y_{n-3} = hf_{n+1},$$

$$k = 5: \quad \frac{137}{60}y_{n+1} - 5y_n + 5y_{n-1} - \frac{10}{3}y_{n-2} + \frac{5}{4}y_{n-3} - \frac{1}{5}y_{n-4} = hf_{n+1},$$

$$k = 6: \quad \frac{147}{60}y_{n+1} - 6y_n + \frac{15}{2}y_{n-1} - \frac{20}{3}y_{n-2} + \frac{15}{4}y_{n-3} - \frac{6}{5}y_{n-4} + \frac{1}{6}y_{n-5} = hf_{n+1}.$$

For $k > 6$ the BDF-methods are unstable (see Section III.3).

Exercises

1. Let the differential equation $y' = y^2$, $y(0) = 1$ and the exact starting values $y_i = 1/(1 - x_i)$ for $i = 0, 1, \dots, k-1$ be given. Apply the methods of Adams and study the expression $y(x_k) - y_k$ for small step sizes.
2. Consider the differential equation at the beginning of this section. It describes the form of a drop and can be written as (F. Bashforth 1883, page 26; the same problem as Exercise 2 of Section II.1 in a different coordinate system)

$$\frac{dx}{d\varphi} = \varrho \cos \varphi, \quad \frac{dz}{d\varphi} = \varrho \sin \varphi \quad (1.24)$$

where

$$\frac{1}{\varrho} + \frac{\sin \varphi}{x} = 2 + \beta z. \quad (1.25)$$

ϱ may be considered as a function of the coordinates x and z . It can be interpreted as the radius of curvature and φ denotes the angle between the normal to the curve and the z -axis (see Fig. 1.7 for $\beta = 3$). The initial values are given by $x(0) = 0$, $z(0) = 0$, $\varrho(0) = 1$.

Solve the above differential equation along the lines of J.C. Adams:

- a) Assuming

$$\varrho = 1 + b_2\varphi^2 + b_4\varphi^4 + \dots$$

and inserting this expression into (1.24) we obtain after integration the truncated Taylor series of $x(\varphi)$ and $z(\varphi)$ in terms of b_2, b_4, \dots . These parameters can then be calculated from (1.25) by comparing the coefficients of φ^m . In this way one obtains the solution for small values of φ (starting values).

- b) Use one of the proposed multistep formulas and calculate the solution for fixed β (say $\beta = 3$) over the interval $[0, \pi]$.

III.2 Local Error and Order Conditions

You know, I am a multistep man . . . and don't tell anybody, but the first program I wrote for the first Swedish computer was a Runge-Kutta code . . .
(G. Dahlquist, 1982, after some glasses of wine; printed with permission)

A general theory of multistep methods was started by the work of Dahlquist (1956, 1959), and became famous through the classical book of Henrici (1962). All multistep formulas considered in the previous section have this in common that the numerical approximations y_i as well as the values f_i appear linearly. We thus consider the general difference equation

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \dots + \alpha_0 y_n = h(\beta_k f_{n+k} + \dots + \beta_0 f_n) \quad (2.1)$$

which includes all considered methods as special cases. In this formula the α_i and β_i are real parameters, h denotes the step size and

$$f_i = f(x_i, y_i), \quad x_i = x_0 + ih.$$

Throughout this chapter we shall assume that

$$\alpha_k \neq 0, \quad |\alpha_0| + |\beta_0| > 0. \quad (2.2)$$

The first assumption expresses the fact that the implicit equation (2.1) can be solved with respect to y_{n+k} at least for sufficiently small h . The second relation in (2.2) can always be achieved by reducing the index k , if necessary.

Formula (2.1) will be called a *linear multistep method* or more precisely a *linear k -step method*. We also distinguish between *explicit* ($\beta_k = 0$) and *implicit* ($\beta_k \neq 0$) multistep methods.

Local Error of a Multistep Method

As the numerical solution of a multistep method does not depend only on the initial value problem (1.1) but also on the choice of the starting values, the definition of the local error is not as straightforward as for one-step methods (compare Sections II.2 and II.3).

Definition 2.1. The *local error* of the multistep method (2.1) is defined by

$$y(x_k) - y_k$$

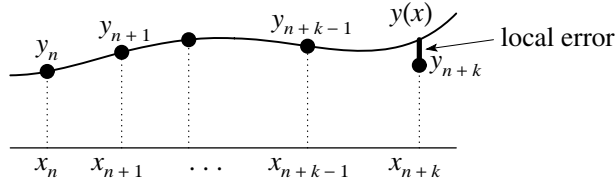


Fig. 2.1. Illustration of the local error

where $y(x)$ is the exact solution of $y' = f(x, y)$, $y(x_0) = y_0$, and y_k is the numerical solution obtained from (2.1) by using the exact starting values $y_i = y(x_i)$ for $i = 0, 1, \dots, k-1$ (see Fig. 2.1).

In the case $k = 1$ this definition coincides with the definition of the local error for one-step methods. In order to show the connection with other possible definitions of the local error, we associate with (2.1) the linear difference operator L defined by

$$L(y, x, h) = \sum_{i=0}^k \left(\alpha_i y(x + ih) - h\beta_i y'(x + ih) \right). \quad (2.3)$$

Here $y(x)$ is some differentiable function defined on an interval that contains the values $x + ih$ for $i = 0, 1, \dots, k$.

Lemma 2.2. Consider the differential equation (1.1) with $f(x, y)$ continuously differentiable and let $y(x)$ be its solution. For the local error one has

$$y(x_k) - y_k = \left(\alpha_k I - h\beta_k \frac{\partial f}{\partial y}(x_k, \eta) \right)^{-1} L(y, x_0, h).$$

Here η is some value between $y(x_k)$ and y_k , if f is a scalar function. In the case of a vector valued function f , the matrix $\frac{\partial f}{\partial y}(x_k, \eta)$ is the Jacobian whose rows are evaluated at possibly different values lying on the segment joining $y(x_k)$ and y_k .

Proof. By Definition 2.1, y_k is determined implicitly by the equation

$$\sum_{i=0}^{k-1} \left(\alpha_i y(x_i) - h\beta_i f(x_i, y(x_i)) \right) + \alpha_k y_k - h\beta_k f(x_k, y_k) = 0.$$

Inserting (2.3) we obtain

$$L(y, x_0, h) = \alpha_k (y(x_k) - y_k) - h\beta_k (f(x_k, y(x_k)) - f(x_k, y_k))$$

and the statement follows from the mean value theorem. \square

This lemma shows that $\alpha_k^{-1}L(y, x_0, h)$ is essentially equal to the local error. Sometimes this term is also called *the* local error (Dahlquist 1956, 1959). For explicit methods both expressions are equal.

Order of a Multistep Method

Once the local error of a multistep method is defined, one can introduce the concept of order in the same way as for one-step methods.

Definition 2.3. The multistep method (2.1) is said to be of *order* p , if one of the following two conditions is satisfied:

- i) for all sufficiently regular functions $y(x)$ we have $L(y, x, h) = \mathcal{O}(h^{p+1})$;
- ii) the local error of (2.1) is $\mathcal{O}(h^{p+1})$ for all sufficiently regular differential equations (1.1).

Observe that by Lemma 2.2 the above conditions (i) and (ii) are equivalent. Our next aim is to characterize the order of a multistep method in terms of the free parameters α_i and β_i . Dahlquist (1956) was the first to observe the fundamental role of the polynomials

$$\begin{aligned}\varrho(\zeta) &= \alpha_k \zeta^k + \alpha_{k-1} \zeta^{k-1} + \dots + \alpha_0 \\ \sigma(\zeta) &= \beta_k \zeta^k + \beta_{k-1} \zeta^{k-1} + \dots + \beta_0.\end{aligned}\tag{2.4}$$

They will be called the *generating polynomials* of the multistep method (2.1).

Theorem 2.4. *The multistep method (2.1) is of order p , if and only if one of the following equivalent conditions is satisfied:*

- i) $\sum_{i=0}^k \alpha_i = 0$ and $\sum_{i=0}^k \alpha_i i^q = q \sum_{i=0}^k \beta_i i^{q-1}$ for $q = 1, \dots, p$;
- ii) $\varrho(e^h) - h\sigma(e^h) = \mathcal{O}(h^{p+1})$ for $h \rightarrow 0$;
- iii) $\frac{\varrho(\zeta)}{\log \zeta} - \sigma(\zeta) = \mathcal{O}((\zeta - 1)^p)$ for $\zeta \rightarrow 1$.

Proof. Expanding $y(x + ih)$ and $y'(x + ih)$ into a Taylor series and inserting these series (truncated if necessary) into (2.3) yields

$$\begin{aligned}L(y, x, h) &= \sum_{i=0}^k \left(\alpha_i \sum_{q \geq 0} \frac{i^q}{q!} h^q y^{(q)}(x) - h \beta_i \sum_{r \geq 0} \frac{i^r}{r!} h^r y^{(r+1)}(x) \right) \\ &= y(x) \sum_{i=0}^k \alpha_i + \sum_{q \geq 1} \frac{h^q}{q!} y^{(q)}(x) \left(\sum_{i=0}^k \alpha_i i^q - q \sum_{i=0}^k \beta_i i^{q-1} \right).\end{aligned}\tag{2.5}$$

This implies the equivalence of condition (i) with $L(y, x, h) = \mathcal{O}(h^{p+1})$ for all sufficiently regular functions $y(x)$.

It remains to prove that the three conditions of Theorem 2.4 are equivalent. The identity

$$L(\exp, 0, h) = \varrho(e^h) - h\sigma(e^h)$$

where \exp denotes the exponential function, together with

$$L(\exp, 0, h) = \sum_{i=0}^k \alpha_i + \sum_{q \geq 1} \frac{h^q}{q!} \left(\sum_{i=0}^k \alpha_i i^q - q \sum_{i=0}^k \beta_i i^{q-1} \right),$$

which follows from (2.5), shows the equivalence of the conditions (i) and (ii).

By use of the transformation $\zeta = e^h$ (or $h = \log \zeta$) condition (ii) can be written in the form

$$\varrho(\zeta) - \log \zeta \cdot \sigma(\zeta) = \mathcal{O}((\log \zeta)^{p+1}) \quad \text{for } \zeta \rightarrow 1.$$

But this condition is equivalent to (iii), since

$$\log \zeta = (\zeta - 1) + \mathcal{O}((\zeta - 1)^2) \quad \text{for } \zeta \rightarrow 1. \quad \square$$

Remark. The conditions for a multistep method to be of order 1, which are usually called *consistency* conditions, can also be written in the form

$$\varrho(1) = 0, \quad \varrho'(1) = \sigma(1). \quad (2.6)$$

Once the proofs of the above order conditions have been understood, it is not difficult to treat the more general situation of non-equidistant grids (see Section III.5 and the book of Stetter (1973), p. 191).

Example 2.5. *Order of the explicit Adams methods.* Let us first investigate for which differential equations the explicit Adams methods give theoretically the exact solution. This is the case if the polynomial $p(t)$ of (1.4) is equal to $f(t, y(t))$. Suppose now that $f(t, y) = f(t)$ does not depend on y and is a polynomial of degree less than k . Then the explicit Adams methods integrate the differential equations

$$y' = qx^{q-1}, \quad \text{for } q = 0, 1, \dots, k$$

exactly. This means that the local error is zero and hence, by Lemma 2.2,

$$0 = L(x^q, 0, h) = h^q \left(\sum_{i=0}^k \alpha_i i^q - q \sum_{i=0}^k \beta_i i^{q-1} \right) \quad \text{for } q = 0, \dots, k.$$

This is just condition (i) of Theorem 2.4 with $p = k$ so that the order of the explicit Adams methods is at least k . In fact it will be shown that the order of these methods is not greater than k (Example 2.7).

Example 2.6. For *implicit Adams methods* the polynomial $p^*(t)$ of (1.8) has degree one higher than that of $p(t)$. Thus the same considerations as in Example 2.5 show that these methods have order at least $k + 1$.

All methods of Section III.1 can be treated analogously (see Exercise 3 and Table 2.1).

Table 2.1. Order and error constant of multistep methods

method	formula	order	error constant
explicitAdams	(1.5)	k	γ_k
implicitAdams	(1.9)	$k + 1$	γ_{k+1}^*
midpoint rule	(1.13')	2	$1/6$
Nyström, $k > 2$	(1.13)	k	$\kappa_k/2$
Milne, $k = 2$	(1.15')	4	$-1/180$
Milne-Simpson, $k > 3$	(1.15)	$k + 1$	$\kappa_{k+1}^*/2$
BDF	(1.22')	k	$-1/(k + 1)$

Error Constant

The order of a multistep method indicates how fast the error tends to zero if $h \rightarrow 0$. Different methods of the *same* order, however, can have different errors; they are distinguished by the *error constant*. Formula (2.5) shows that the difference operator L , associated with a p th order multistep method, is such that for all sufficiently regular functions $y(x)$

$$L(y, x, h) = C_{p+1} h^{p+1} y^{(p+1)}(x) + \mathcal{O}(h^{p+2}) \quad (2.7)$$

where the constant C_{p+1} is given by

$$C_{p+1} = \frac{1}{(p+1)!} \left(\sum_{i=0}^k \alpha_i i^{p+1} - (p+1) \sum_{i=0}^k \beta_i i^p \right). \quad (2.8)$$

This constant is not suitable as a measure of accuracy, since multiplication of formula (2.1) by a constant can give any value for C_{p+1} , whereas the numerical solution $\{y_n\}$ remains unchanged. A better choice would be the constant $\alpha_k^{-1} C_{p+1}$, since the local error of a multistep method is given by (Lemma 2.2 and formula (2.7))

$$y(x_k) - y_k = \alpha_k^{-1} C_{p+1} h^{p+1} y^{(p+1)}(x_0) + \mathcal{O}(h^{p+2}). \quad (2.9)$$

For several reasons, however, this is not yet a satisfactory definition, as we shall see from the following motivation: let

$$e_n = \frac{y(x_n) - y_n}{h^p}$$

be the global error scaled by h^p , and assume for this motivation that $e_n = \mathcal{O}(1)$. Subtracting (2.1) from (2.3) and using (2.7) we have

$$\begin{aligned} \sum_{i=0}^k \alpha_i e_{n+i} &= h^{1-p} \sum_{i=0}^k \beta_i \left(f(x_{n+i}, y(x_{n+i})) - f(x_{n+i}, y_{n+i}) \right) \\ &\quad + C_{p+1} h y^{(p+1)}(x_n) + \mathcal{O}(h^2). \end{aligned} \quad (2.10)$$

The point is now to use

$$y^{(p+1)}(x_n) = \frac{1}{\sigma(1)} \sum_{i=0}^k \beta_i y^{(p+1)}(x_{n+i}) + \mathcal{O}(h) \quad (2.11)$$

which brings the error term in (2.10) inside the sum with the β_i . We linearize

$$f(x_{n+i}, y(x_{n+i})) - f(x_{n+i}, y_{n+i}) = \frac{\partial f}{\partial y}(x_{n+i}, y(x_{n+i})) h^p e_{n+i} + \mathcal{O}(h^{2p})$$

and insert this together with (2.11) into (2.10). Neglecting the $\mathcal{O}(h^2)$ and $\mathcal{O}(h^{2p})$ terms, we can interpret the obtained formula as the multistep method applied to

$$e'(x) = \frac{\partial f}{\partial y}(x, y(x)) e(x) + C y^{(p+1)}(x), \quad e(x_0) = 0, \quad (2.12)$$

where

$$C = \frac{C_{p+1}}{\sigma(1)} \quad (2.13)$$

is seen to be a natural measure for the global error and is therefore called *the error constant*.

Another derivation of Definition (2.13) will be given in the section on global convergence (see Exercise 2 of Section III.4). Further, the solution of (2.12) gives the first term of the asymptotic expansion of the global error (see Section III.9).

Example 2.7. *Error constant of the explicit Adams methods.* Consider the differential equation $y' = f(x)$ with $f(x) = (k+1)x^k$, the exact solution of which is $y(x) = x^{k+1}$. As this differential equation is integrated exactly by the $(k+1)$ -step explicit Adams method (see Example 2.5), we have

$$y(x_k) - y(x_{k-1}) = h \sum_{j=0}^k \gamma_j \nabla^j f_{k-1}.$$

The local error of the k -step explicit Adams method (1.5) is therefore given by

$$y(x_k) - y_k = h \gamma_k \nabla^k f_{k-1} = h^{k+1} \gamma_k f^{(k)}(x_0) = h^{k+1} \gamma_k y^{(k+1)}(x_0).$$

As $\gamma_k \neq 0$, this formula shows that the order of the k -step method is not greater than k (compare Example 2.5). Furthermore, since $\alpha_k = 1$, a comparison with formula (2.9) yields $C_{k+1} = \gamma_k$. Finally, for Adams methods we have $\varrho(\zeta) = \zeta^k - \zeta^{k-1}$ and $\varrho'(1) = 1$, so that by the use of (2.6) the error constant is given by $C = \gamma_k$.

The error constants of all other previously considered multistep methods are summarized in Table 2.1 (observe that $\sigma(1) = 2$ for explicit Nyström and Milne-Simpson methods).

Irreducible Methods

Let $\varrho(\zeta)$ and $\sigma(\zeta)$ of formula (2.4) be the generating polynomials of (2.1) and suppose that they have a common factor $\varphi(\zeta)$. Then the polynomials

$$\varrho^*(\zeta) = \frac{\varrho(\zeta)}{\varphi(\zeta)}, \quad \sigma^*(\zeta) = \frac{\sigma(\zeta)}{\varphi(\zeta)},$$

are the generating polynomials of a new and simpler multistep method. Using the shift operator E , defined by

$$Ey_n = y_{n+1} \quad \text{or} \quad Ey(x) = y(x+h),$$

this multistep method can be written in compact form as

$$\varrho^*(E)y_n = h\sigma^*(E)f_n.$$

Multiplication by $\varphi(E)$ shows that any solution $\{y_n\}$ of this method is also a solution of $\varrho(E)y_n = h\sigma(E)f_n$. The two methods are thus essentially equal. Denote by L^* the difference operator associated with the new reduced method, and by C_{p+1}^* the constant given by (2.7). As

$$\begin{aligned} L(y, x, h) &= \varphi(E)L^*(y, x, h) = C_{p+1}^* h^{p+1} \varphi(E)y^{(p+1)}(x) + \mathcal{O}(h^{p+2}) \\ &= C_{p+1}^* \varphi(1) h^{p+1} y^{(p+1)}(x) + \mathcal{O}(h^{p+2}) \end{aligned}$$

one immediately obtains $C_{p+1} = \varphi(1)C_{p+1}^*$ and therefore also the relation

$$C_{p+1}/\sigma(1) = C_{p+1}^*/\sigma^*(1)$$

holds. Both methods thus have the same error constant.

The above analysis has shown that multistep methods whose generating polynomials have a common factor are not interesting. We therefore usually assume that

$$\varrho(\zeta) \text{ and } \sigma(\zeta) \text{ have no common factor.} \quad (2.14)$$

Multistep methods satisfying this property are called *irreducible*.

The Peano Kernel of a Multistep Method

The order and the error constant above do not yet give a complete description of the error, since the subsequent terms of the series for the error may be much larger than C_{p+1} . Several attempts have therefore been made, originally for the error of a quadrature formula, to obtain a complete description of the error. The following discussion is an extension of the ideas of Peano (1913).

Theorem 2.8. *Let the multistep method (2.1) be of order p and let q ($1 \leq q \leq p$) be an integer. For any $(q+1)$ -times continuously differentiable function $y(x)$ we then have*

$$L(y, x, h) = h^{q+1} \int_0^k K_q(s) y^{(q+1)}(x + sh) ds, \quad (2.15)$$

where

$$K_q(s) = \frac{1}{q!} \sum_{i=0}^k \alpha_i (i-s)_+^q - \frac{1}{(q-1)!} \sum_{i=0}^k \beta_i (i-s)_+^{q-1} \quad (2.16a)$$

with

$$(i-s)_+^r = \begin{cases} (i-s)^r & \text{for } i-s > 0 \\ 0 & \text{for } i-s \leq 0. \end{cases}$$

$K_q(s)$ is called the q th Peano kernel of the multistep method (2.1).

Remark. We see from (2.16a) that $K_q(s)$ is a piecewise polynomial and satisfies

$$K_q(s) = \frac{1}{q!} \sum_{i=j}^k \alpha_i (i-s)^q - \frac{1}{(q-1)!} \sum_{i=j}^k \beta_i (i-s)^{q-1} \quad \text{for } s \in [j-1, j). \quad (2.16b)$$

Proof. Taylor's theorem with the integral representation of the remainder yields

$$\begin{aligned} y(x+ih) &= \sum_{r=0}^q \frac{i^r}{r!} h^r y^{(r)}(x) + h^{q+1} \int_0^i \frac{(i-s)^q}{q!} y^{(q+1)}(x+sh) ds, \\ hy'(x+ih) &= \sum_{r=1}^q \frac{i^{r-1}}{(r-1)!} h^r y^{(r)}(x) + h^{q+1} \int_0^i \frac{(i-s)^{q-1}}{(q-1)!} y^{(q+1)}(x+sh) ds. \end{aligned}$$

Inserting these two expressions into (2.3), the same considerations as in the proof of Theorem 2.4 show that for $q \leq p$ the polynomials before the integral cancel. The statement then follows from

$$\int_0^i \frac{(i-s)^q}{q!} y^{(q+1)}(x+sh) ds = \int_0^k \frac{(i-s)_+^q}{q!} y^{(q+1)}(x+sh) ds. \quad \square$$

Besides the representation (2.16), the Peano kernel $K_q(s)$ has the following properties:

$$K_q(s) = 0 \text{ for } s \in (-\infty, 0) \cup [k, \infty) \text{ and } q = 1, \dots, p; \quad (2.17)$$

$$\begin{aligned} K_q(s) &\text{ is } (q-2)\text{-times continuously differentiable and} \\ K_q'(s) &= -K_{q-1}(s) \text{ for } q = 2, \dots, p \text{ (for } q = 2 \text{ piecewise);} \end{aligned} \quad (2.18)$$

$$\begin{aligned} K_1(s) &\text{ is a piecewise linear function with discontinuities at } \\ &0, 1, \dots, k. \text{ It has a jump of size } \beta_j \text{ at the point } j \text{ and its} \\ &\text{slope over the interval } (j-1, j) \text{ is given by } -(\alpha_j + \alpha_{j+1} + \\ &\dots + \alpha_k); \end{aligned} \quad (2.19)$$

$$\text{For the constant } C_{p+1} \text{ of (2.8) we have } C_{p+1} = \int_0^k K_p(s) ds. \quad (2.20)$$

The proofs of Statements (2.17) to (2.20) are as follows: it is an immediate consequence of the definition of the Peano kernel that $K_q(s) = 0$ for $s \geq k$ and $q \leq p$. In order to prove that $K_q(s) = 0$ also for $s < 0$ we consider the polynomial $y(x) = (x-s)^q$ with s as a parameter. Theorem 2.8 then shows that

$$L(y, 0, 1) = \sum_{i=0}^k \alpha_i (i-s)^q - q \sum_{i=0}^k \beta_i (i-s)^{q-1} \equiv 0 \quad \text{for } q \leq p$$

and hence $K_q(s) = 0$ for $s < 0$. This gives (2.17). The relation (2.18) is seen by partial integration of (2.15). As an example, the Peano kernels for the 3-step Nyström method (1.13'') are plotted in Fig. 2.2.

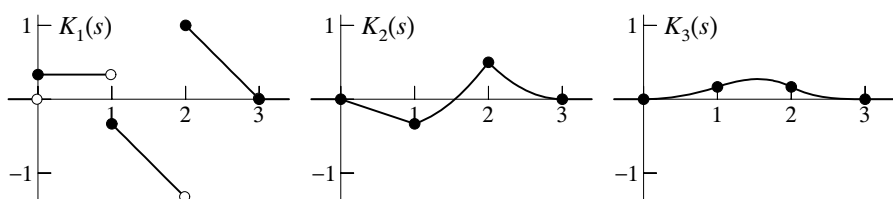


Fig. 2.2. Peano kernels of the 3-step Nyström method

Exercises

1. Construction of multistep methods. Let $\varrho(\zeta)$ be a k th degree polynomial satisfying $\varrho(1) = 0$.
 - a) There exists exactly one polynomial $\sigma(\zeta)$ of degree $\leq k$, such that the order of the corresponding multistep method is at least $k + 1$.
 - b) There exists exactly one polynomial $\sigma(\zeta)$ of degree $< k$, such that the corresponding multistep method, which is then explicit, has order at least k .

Hint. Use condition (iii) of Theorem 2.4.

2. Find the multistep method of the form

$$y_{n+2} + \alpha_1 y_{n+1} + \alpha_0 y_n = h(\beta_1 f_{n+1} + \beta_0 f_n)$$

of the highest possible order. Apply this formula to the example $y' = y$, $y(0) = 1$, $h = 0.1$.

3. Verify that the order and the error constant of the BDF-formulas are those of Table 2.1.
4. Show that the Peano kernel $K_p(s)$ does not change sign for the explicit and implicit Adams methods, nor for the BDF-formulas. Deduce from this property that

$$L(y, x, h) = h^{p+1} C_{p+1} y^{(p+1)}(\zeta) \quad \text{with } \zeta \in (x, x + kh)$$

where the constant C_{p+1} is given by (2.8).

5. Let $y(x)$ be an exact solution of $y' = f(x, y)$ and let $y_i = y(x_i)$, $i = 0, 1, \dots, k - 1$. Assume that f is continuous and satisfies a Lipschitz condition with respect to y (f not necessarily differentiable). Prove that for consistent multistep methods (i.e., methods with (2.6)) the local error satisfies

$$\|y(x_k) - y_k\| \leq h\omega(h)$$

where $\omega(h) \rightarrow 0$ for $h \rightarrow 0$.

III.3 Stability and the First Dahlquist Barrier

... hat der Verfasser seither öfters Verfahren zur numerischen Integration von Differentialgleichungen beobachtet, die, obschon zwar mit bestechend kleinem Abbruchfehler behaftet, doch die grosse Gefahr der numerischen Instabilität in sich bergen.

(H. Rutishauser 1952)

Rutishauser observed in his famous paper that high order and a small local error are not sufficient for a useful multistep method. The numerical solution can be “unstable”, even though the step size h is taken very small. The same observation was made by Todd (1950), who applied certain difference methods to second order differential equations. Our presentation will mainly follow the lines of Dahlquist (1956), where this effect has been studied systematically. An interesting presentation of the historical development of numerical stability concepts can be found in Dahlquist (1985) “33 years of numerical instability, Part I”.

Let us start with an example, taken from Dahlquist (1956). Among all explicit 2-step methods we consider the formula with the highest order (see Exercise 2 of Section III.2). A short calculation using Theorem 2.4 shows that this method of order 3 is given by

$$y_{n+2} + 4y_{n+1} - 5y_n = h(4f_{n+1} + 2f_n). \quad (3.1)$$

Application to the differential equation

$$y' = y, \quad y(0) = 1 \quad (3.2)$$

yields the linear difference relation

$$y_{n+2} + 4(1-h)y_{n+1} - (5+2h)y_n = 0. \quad (3.3)$$

As starting values we take $y_0 = 1$ and $y_1 = \exp(h)$, the values on the exact solution. The numerical solution together with the exact solution $\exp(x)$ is plotted in Fig. 3.1 for the step sizes $h = 1/10$, $h = 1/20$, $h = 1/40$, etc. In spite of the small local error, the results are very bad and become even worse as the step size decreases.

An explanation for this effect can easily be given. As usual for linear difference equations (Dan. Bernoulli 1728, Lagrange 1775), we insert $y_j = \zeta^j$ into (3.3). This leads to the characteristic equation

$$\zeta^2 + 4(1-h)\zeta - (5+2h) = 0. \quad (3.4)$$

The general solution of (3.3) is then given by

$$y_n = A\zeta_1^n(h) + B\zeta_2^n(h) \quad (3.5)$$

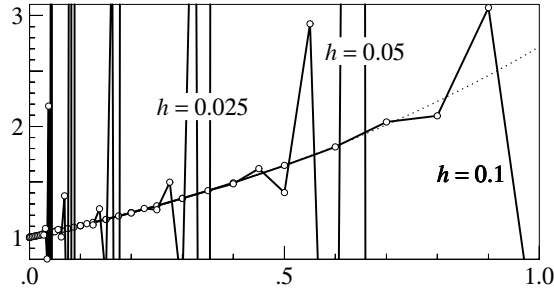


Fig. 3.1. Numerical solution of the unstable method (3.1)

where

$$\zeta_1(h) = 1 + h + \mathcal{O}(h^2), \quad \zeta_2(h) = -5 + \mathcal{O}(h)$$

are the roots of (3.4) and the coefficients A and B are determined by the starting values y_0 and y_1 . Since $\zeta_1(h)$ approximates $\exp(h)$, the first term in (3.5) approximates the exact solution $\exp(x)$ at the point $x = nh$. The second term in (3.5), often called a *parasitic solution*, is the one which causes trouble in our method: since for $h \rightarrow 0$ the absolute value of $\zeta_2(h)$ is larger than one, this parasitic solution becomes very large and dominates the solution y_n for increasing n .

We now turn to the stability discussion of the general method (2.1). The essential part is the behaviour of the solution as $n \rightarrow \infty$ (or $h \rightarrow 0$) with nh fixed. We see from (3.3) that for $h \rightarrow 0$ we obtain

$$\alpha_k y_{n+k} + \alpha_{k-1} y_{n+k-1} + \dots + \alpha_0 y_n = 0. \quad (3.6)$$

This can be interpreted as the numerical solution of the method (2.1) for the differential equation

$$y' = 0. \quad (3.7)$$

We put $y_j = \zeta^j$ in (3.6), divide by ζ^n , and find that ζ must be a root of

$$\varrho(\zeta) = \alpha_k \zeta^k + \alpha_{k-1} \zeta^{k-1} + \dots + \alpha_0 = 0. \quad (3.8)$$

As in Section I.13, we again have some difficulty when (3.8) possesses a root of multiplicity $m > 1$. In this case (Lagrange 1792, see Exercise 1 below) $y_n = n^{j-1} \zeta^n$ ($j = 1, \dots, m$) are solutions of (3.6) and we obtain by superposition:

Lemma 3.1. *Let ζ_1, \dots, ζ_l be the roots of $\varrho(\zeta)$, of respective multiplicity m_1, \dots, m_l . Then the general solution of (3.6) is given by*

$$y_n = p_1(n) \zeta_1^n + \dots + p_l(n) \zeta_l^n \quad (3.9)$$

where the $p_j(n)$ are polynomials of degree $m_j - 1$. □

Formula (3.9) shows us that for boundedness of y_n , as $n \rightarrow \infty$, we need that the roots of (3.8) lie in the unit disc and that the roots on the unit circle be simple.

Definition 3.2. The multistep method (2.1) is called *stable*, if the generating polynomial $\varrho(\zeta)$ (formula (3.8)) satisfies the *root condition*, i.e.,

- i) The roots of $\varrho(\zeta)$ lie on or within the unit circle;
- ii) The roots on the unit circle are simple.

Remark. In order to distinguish this stability concept from others, it is sometimes called *zero-stability* or, in honour of Dahlquist, also *D-stability*.

Examples. For the explicit and implicit *Adams methods*, $\varrho(\zeta) = \zeta^k - \zeta^{k-1}$. Besides the simple root 1, there is a $(k-1)$ -fold root at 0. The Adams methods are therefore stable.

The same is true for the explicit *Nyström* and the *Milne-Simpson methods*, where $\varrho(\zeta) = \zeta^k - \zeta^{k-2}$. Note that here we have a simple root at -1 . This root can be dangerous for certain differential equations (see Section III.9 and Section V.1 of Volume II).

Stability of the BDF-Formulas

The investigation of the stability of the BDF-formulas is more difficult. As the characteristic polynomial of $\nabla^j y_{k+n} = 0$ is given by $\zeta^{k-j}(\zeta-1)^j = 0$ it follows from the representation (1.22') that the generating polynomial $\varrho(\zeta)$ of the BDF-formulas has the form

$$\varrho(\zeta) = \sum_{j=1}^k \frac{1}{j} \zeta^{k-j} (\zeta-1)^j. \quad (3.10)$$

In order to study the zeros of (3.10) it is more convenient to consider the polynomial

$$p(z) = (1-z)^k \varrho\left(\frac{1}{1-z}\right) = \sum_{j=1}^k \frac{z^j}{j} \quad (3.11)$$

via the transformation $\zeta = 1/(1-z)$. This polynomial is just the k th partial sum of $-\log(1-z)$. As the roots of $p(z)$ and $\varrho(\zeta)$ are related by the above transformation, we have:

Lemma 3.3. *The k -step BDF-formula (1.22') is stable iff all roots of the polynomial (3.11) are outside the disc $\{z; |z-1| \leq 1\}$, with simple roots allowed on the boundary.* \square

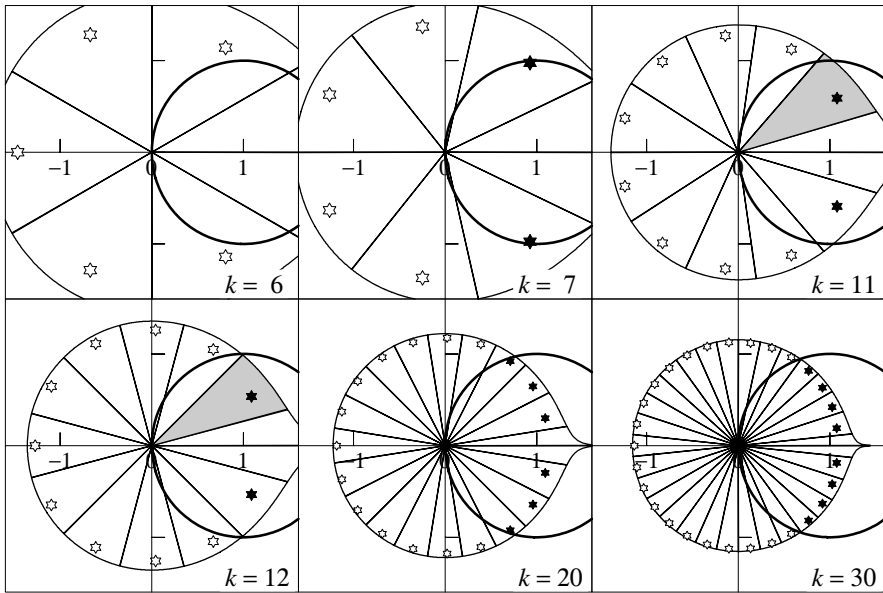


Fig. 3.2. Roots of the polynomial $p(z)$ of (3.11)

The roots of (3.11) are displayed in Fig. 3.2 for different values of k .

Theorem 3.4. *The k -step BDF-formula (1.22') is stable for $k \leq 6$, and unstable for $k \geq 7$.*

Proof. The first assertion can be verified simply by a finite number of numerical calculations (see Fig. 3.2). This was first observed by Mitchell & Craggs (1953). The second statement, however, contains an infinity of cases and is more difficult. The first complete proof was given by Cryer (1971) in a technical report, a condensed version of which is published in Cryer (1972). A second proof is given in Creedon & Miller (1975) (see also Grigorieff (1977), p. 135), based on the Schur-Cohn criterion. This proof is outlined in Exercise 4 below. The following proof, which is given in Hairer & Wanner (1983), is based on the representation

$$p(z) = \int_0^z \sum_{j=1}^k \zeta^{j-1} d\zeta = \int_0^z \frac{1-\zeta^k}{1-\zeta} d\zeta = \int_0^r (1 - e^{ik\theta} s^k) \varphi(s) ds \quad (3.12)$$

with

$$\zeta = se^{i\theta}, \quad z = re^{i\theta}, \quad \varphi(s) = \frac{e^{i\theta}}{1 - se^{i\theta}}.$$

We cut the complex plane into k sectors

$$S_j = \left\{ z; \frac{2\pi}{k} \left(j - \frac{1}{2} \right) < \arg(z) < \frac{2\pi}{k} \left(j + \frac{1}{2} \right) \right\}, \quad j = 0, 1, \dots, k-1.$$

On the rays bounding S_j we have $e^{ik\theta} = -1$, so that from (3.12)

$$p(z) = \int_0^r (1 + s^k) \varphi(s) ds$$

with a *positive* weight function. Therefore, $p(z)$ always lies in the sector between $e^{i\theta}$ and $e^{i\pi} = -1$, which contains all values $\varphi(s)$ (see Theorem 1.1 on page 1 of Marden (1966)). So no revolution of $\arg(p(z))$ is possible on these rays, and due to the one revolution of $\arg(z^k)$ at infinity between $\theta = 2\pi(j - 1/2)/k$ and $\theta = 2\pi(j + 1/2)/k$ the principle of the argument (e.g., Henrici (1974), p. 278) implies (see Fig. 3.3) that in each sector S_j ($j = 1, \dots, k - 1$, with the exception of $j = 0$) there lies exactly one root of $p(z)$.

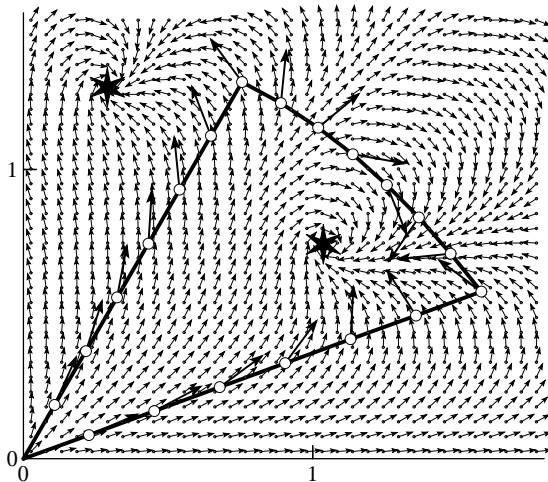


Fig. 3.3. Argument of $p(z)$ of (3.11)

In order to complete the proof, we still have to bound the zeros of $p(z)$ from above: we observe that in (3.12) the term s^k becomes large for $s > 1$. We therefore partition (3.12) into two integrals $p(z) = I_1 - I_2$, where

$$I_1 = \int_0^r \varphi(s) ds - \int_0^1 e^{ik\theta} s^k \varphi(s) ds, \quad I_2 = e^{ik\theta} \int_1^r s^k \varphi(s) ds.$$

Since $|\varphi(s)| \leq B(\theta)$ where

$$B(\theta) = \begin{cases} |\sin \theta|^{-1} & \text{if } 0 < \theta \leq \pi/2 \text{ or } 3\pi/2 \leq \theta < 2\pi, \\ 1 & \text{otherwise,} \end{cases}$$

we obtain

$$|I_1| \leq \left(r + \frac{1}{k+1}\right) B(\theta) < r B(\theta) \frac{k+2}{k+1}, \quad (r > 1). \quad (3.13)$$

Secondly, since s^k is positive,

$$I_2 = e^{ik\theta} \Phi \int_1^r s^k ds \quad \text{with} \quad \Phi \in \text{convex hull of } \{\varphi(s); 1 \leq s \leq r\}.$$

Any element of the above convex hull can be written in the form

$$\Phi = \alpha \varphi(s_1) + (1 - \alpha) \varphi(s_2) = \frac{\varphi(s_1) \varphi(s_2)}{\varphi(\hat{s})}$$

with $\hat{s} = \alpha s_2 + (1 - \alpha) s_1$, $0 \leq \alpha \leq 1$, $1 \leq s_1, s_2 \leq r$. Since $|\varphi(s)|$ decreases monotonically for $s \geq 1$, we have $|\Phi| \geq |\varphi(r)|$. Some elementary geometry then leads to $|\Phi| \geq 1/2r$ and we get

$$|I_2| \geq \frac{r^{k+1} - 1}{2r(k+1)} > \frac{r(r^{k-1} - 1)}{2k+2}, \quad (r > 1). \quad (3.14)$$

From (3.13) and (3.14) we see that

$$r \geq R(\theta) = ((2k+4)B(\theta) + 1)^{1/(k-1)} \quad (3.15)$$

implies $|I_2| > |I_1|$, so that $p(z)$ cannot be zero. The curve $R(\theta)$ is also plotted in Fig. 3.2 and cuts from the sectors S_j what we call Madame Imhof's cheese pie, each slice of which (with $j \neq 0$) must contain precisely one zero of $p(z)$. A simple analysis shows that for $k = 12$ the cheese pie, cut from S_1 , is small enough to ensure the presence of zeros of $p(z)$ inside the disc $\{z; |z - 1| \leq 1\}$. As $R(\theta)$, for fixed θ , as well as $R(\pi/k)$ are monotonically decreasing in k , the same is true for all $k \geq 12$.

For $6 < k < 12$ numerical calculations show that the method is unstable (see Fig. 3.2 or Exercise 4). \square

Highest Attainable Order of Stable Multistep Methods

It is a natural task to investigate the stability of the multistep methods with highest possible order. This has been performed by Dahlquist (1956), resulting in the famous "first Dahlquist-barrier".

Counting the order conditions (Theorem 2.4) shows that for order p the parameters of a linear multistep method have to satisfy $p + 1$ linear equations. As $2k + 1$ free parameters are involved (without loss of generality one can assume $\alpha_k = 1$), this suggests that $2k$ is the highest attainable order. Indeed, this can be verified (see Exercise 5). However, these methods are of no practical significance, because we shall prove

Theorem 3.5 (The first Dahlquist-barrier). *The order p of a stable linear k -step method satisfies*

$$\begin{aligned} p &\leq k+2 && \text{if } k \text{ is even,} \\ p &\leq k+1 && \text{if } k \text{ is odd,} \\ p &\leq k && \text{if } \beta_k/\alpha_k \leq 0 \text{ (in particular if the method is explicit).} \end{aligned}$$

We postpone the verification of this theorem and give some notations and lemmas, which will be useful for the proof. First of all we introduce the “Greek-Roman transformation”

$$\zeta = \frac{z+1}{z-1} \quad \text{or} \quad z = \frac{\zeta+1}{\zeta-1}. \quad (3.16)$$

This transformation maps the disk $|\zeta| < 1$ onto the half-plane $\operatorname{Re} z < 0$, the upper half-plane $\operatorname{Im} z > 0$ onto the lower half-plane, the circle $|\zeta| = 1$ to the imaginary axis, the point $\zeta = 1$ to $z = \infty$ and the point $\zeta = -1$ to $z = 0$. We then consider the polynomials

$$\begin{aligned} R(z) &= \left(\frac{z-1}{2}\right)^k \varrho(\zeta) = \sum_{j=0}^k a_j z^j, \\ S(z) &= \left(\frac{z-1}{2}\right)^k \sigma(\zeta) = \sum_{j=0}^k b_j z^j. \end{aligned} \quad (3.17)$$

Since the zeros of $R(z)$ and of $\varrho(\zeta)$ are connected via the transformation (3.16), the stability condition of a multistep method can be formulated in terms of $R(z)$ as follows: all zeros of $R(z)$ lie in the negative half-plane $\operatorname{Re} z \leq 0$ and no multiple zero of $R(z)$ lies on the imaginary axis.

Lemma 3.6. *Suppose the multistep method to be stable and of order at least 0. We then have*

- i) $a_k = 0$ and $a_{k-1} = 2^{1-k} \varrho'(1) \neq 0$;
- ii) *All non-vanishing coefficients of $R(z)$ have the same sign.*

Proof. Dividing formula (3.17) by z^k and putting $z = \infty$, one sees that $a_k = 2^{-k} \varrho(1)$. This expression must vanish, because the method is of order 0. In the same way one gets $a_{k-1} = 2^{1-k} \varrho'(1)$, which is different from zero, since by stability 1 cannot be a multiple root of $\varrho(\zeta)$. The second statement follows from the factorization

$$R(z) = a_{k-1} \prod (z + x_j) \prod ((z + u_j)^2 + v_j^2).$$

where $-x_j$ are the real roots and $-u_j \pm iv_j$ are the conjugate pairs of complex roots. By stability $x_j \geq 0$ and $u_j \geq 0$, implying that all coefficients of $R(z)$ have the same sign. \square

We next express the order conditions of Theorem 2.4 in terms of the polynomials $R(z)$ and $S(z)$.

Lemma 3.7. *The multistep method is of order p if and only if*

$$R(z) \left(\log \frac{z+1}{z-1} \right)^{-1} - S(z) = C_{p+1} \left(\frac{2}{z} \right)^{p-k} + \mathcal{O} \left(\left(\frac{2}{z} \right)^{p-k+1} \right) \quad \text{for } z \rightarrow \infty \quad (3.18)$$

Proof. First, observe that the $\mathcal{O}((\zeta-1)^p)$ term in condition (iii) of Theorem 2.4 is equal to $C_{p+1}(\zeta-1)^p + \mathcal{O}((\zeta-1)^{p+1})$ by formula (2.7). Application of the transformation (3.16) then yields (3.18), because $(\zeta-1) = 2/(z-1) = 2/z + \mathcal{O}((2/z)^2)$ for $z \rightarrow \infty$. \square

Lemma 3.8. *The coefficients of the Laurent series*

$$\left(\log \frac{z+1}{z-1} \right)^{-1} = \frac{z}{2} - \mu_1 z^{-1} - \mu_3 z^{-3} - \mu_5 z^{-5} - \dots \quad (3.19)$$

satisfy $\mu_{2j+1} > 0$ for all $j \geq 0$.

Proof. We consider the branch of $\log \zeta$ which is analytic in the complex ζ -plane cut along the negative real axis and satisfies $\log 1 = 0$. The transformation (3.16) maps this cut onto the segment from -1 to $+1$ on the real axis. The function $\log((z+1)/(z-1))$ is thus analytic on the complex z -plane cut along this segment (see Fig. 3.4). From the formula

$$\log \frac{z+1}{z-1} = \frac{2}{z} \left(1 + \frac{z^{-2}}{3} + \frac{z^{-4}}{5} + \frac{z^{-6}}{7} + \dots \right), \quad (3.20)$$

the existence of (3.19) becomes clear. In order to prove the positivity of the coefficients, we use Cauchy's formula for the coefficients of the function $f(z) = \sum_{n \in \mathbb{Z}} a_n (z - z_0)^n$,

$$a_n = \frac{1}{2\pi i} \int_{\gamma} \frac{f(z)}{(z - z_0)^{n+1}} dz,$$

i.e., in our situation

$$\mu_{2j+1} = -\frac{1}{2\pi i} \int_{\gamma} z^{2j} \left(\log \frac{z+1}{z-1} \right)^{-1} dz$$

(Cauchy 1831; see also Behnke & Sommer 1962). Here γ is an arbitrary curve enclosing the segment $(-1, 1)$, e.g., the curve plotted in Fig. 3.4.

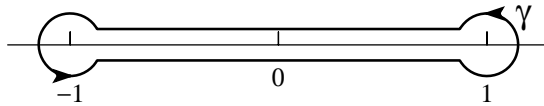


Fig. 3.4. Cut z -plane with curve γ

Observing that $\log((z+1)/(z-1)) = \log((1+x)/(1-x)) - i\pi$ when z approaches the real value $x \in (-1, 1)$ from above, and that $\log((z+1)/(z-1)) = \log((1+x)/(1-x)) + i\pi$ when z approaches x from below, we obtain

$$\begin{aligned}\mu_{2j+1} &= -\frac{1}{2\pi i} \int_{-1}^1 x^{2j} \left[\left(\log \frac{1+x}{1-x} + i\pi \right)^{-1} - \left(\log \frac{1+x}{1-x} - i\pi \right)^{-1} \right] dx \\ &= \int_{-1}^1 x^{2j} \left[\left(\log \frac{1+x}{1-x} \right)^2 + \pi^2 \right]^{-1} dx > 0.\end{aligned}\quad \square$$

For another proof of this lemma, which avoids complex analysis, see Exercise 10.

Proof of Theorem 3.5. We insert the series (3.19) into (3.18) and obtain

$$R(z) \left(\log \frac{z+1}{z-1} \right)^{-1} - S(z) = \text{polynomial}(z) + d_1 z^{-1} + d_2 z^{-2} + \mathcal{O}(z^{-3}) \quad (3.21)$$

where

$$\begin{aligned}d_1 &= -\mu_1 a_0 - \mu_3 a_2 - \mu_5 a_4 - \dots \\ d_2 &= -\mu_3 a_1 - \mu_5 a_3 - \mu_7 a_5 - \dots\end{aligned}\quad (3.22)$$

Lemma 3.6 together with the positivity of the μ_j (Lemma 3.8) implies that all summands in the above formulas for d_1 and d_2 have the same sign. Since $a_{k-1} \neq 0$ we therefore have $d_2 \neq 0$ for k even and $d_1 \neq 0$ for k odd. The first two bounds of Theorem 3.5 are now an immediate consequence of formula (3.18).

Finally, we prove that $p \leq k$ for $\beta_k/\alpha_k \leq 0$: assume, by contradiction, that the order is greater than k . Then by formula (3.18), $S(z)$ is equal to the principal part of $R(z)(\log((z+1)/(z-1)))^{-1}$, and we may write (putting $\mu_j = 0$ for even j)

$$S(z) = R(z) \left(\frac{z}{2} - \sum_{j=1}^{k-1} \mu_j z^{-j} \right) + \sum_{j=1}^{k-1} \left(\sum_{s=j}^{k-1} \mu_s a_{s-j} \right) z^{-j}.$$

Setting $z = 1$ we obtain

$$\frac{S(1)}{R(1)} = \left(\frac{1}{2} - \sum_{j=1}^{k-1} \mu_j \right) + \sum_{j=1}^{k-1} \left(\sum_{s=j}^{k-1} \mu_s a_{s-j} \right) \frac{1}{R(1)}. \quad (3.23)$$

Since by formula (3.17), $S(1) = \beta_k$ and $R(1) = \alpha_k$, it is sufficient to prove $S(1)/R(1) > 0$. Formula (3.19), for $z \rightarrow 1$, gives

$$\sum_{j=1}^{\infty} \mu_j = \frac{1}{2},$$

so that the first summand in (3.23) is strictly positive. The non-negativeness of the second summand is seen from Lemmas 3.6 and 3.8. \square

The stable multistep methods which attain the highest possible order $k + 2$ have a very special structure.

Theorem 3.9. *Stable multistep methods of order $k + 2$ are symmetric, i.e.,*

$$\alpha_j = -\alpha_{k-j}, \quad \beta_j = \beta_{k-j} \quad \text{for all } j. \quad (3.24)$$

Remark. For symmetric multistep methods we have $\varrho(\zeta) = -\zeta^k \varrho(1/\zeta)$ by definition. Since with ζ_i also $1/\zeta_i$ is a zero of $\varrho(\zeta)$, all roots of stable symmetric multistep methods lie on the unit circle and are simple.

Proof. A comparison of the formulas (3.18) and (3.21) shows that $d_1 = 0$ is necessary for order $k + 2$. Since the method is assumed to be stable, Lemma 3.6 implies that all even coefficients of $R(z)$ vanish. Hence, k is even and $R(z)$ satisfies the relation $R(z) = -R(-z)$. By definition of $R(z)$ this relation is equivalent to $\varrho(\zeta) = -\zeta^k \varrho(1/\zeta)$, which implies the first condition of (3.24). Using the above relation for $R(z)$ one obtains from formula (3.18) that $S(z) - S(-z) = \mathcal{O}((2/z)^2)$, implying $S(z) = S(-z)$. If this relation is transformed into an equivalent one for $\sigma(\zeta)$, one gets the second condition of (3.24). \square

Exercises

1. Consider the linear difference equation (3.6) with

$$\varrho(\zeta) = \alpha_k \zeta^k + \alpha_{k-1} \zeta^{k-1} + \dots + \alpha_0$$

as characteristic polynomial. Let ζ_1, \dots, ζ_l be the different roots of $\varrho(\zeta)$ and let $m_j \geq 1$ be the multiplicity of the root ζ_j . Show that for $1 \leq j \leq l$ and $0 \leq i \leq m_j - 1$ the sequences

$$\left\{ \binom{n}{i} \zeta_j^{n-i} \right\}_{n \geq 0}$$

form a system of k linearly independent solutions of (3.6).

2. Show that all roots of the polynomial $p(z)$ of formula (3.11) except the simple root 0 lie in the annulus

$$\frac{k}{k-1} \leq |z| \leq 2.$$

Hint. Use the following lemma, which can be found in Marden (1966), p.137: if all coefficients of the polynomial $a_k z^k + a_{k-1} z^{k-1} + \dots + a_0$ are real and positive, then its roots lie in the annulus $\varrho_1 \leq |z| \leq \varrho_2$ with $\varrho_1 = \min(a_j/a_{j+1})$ and $\varrho_2 = \max(a_j/a_{j+1})$.

3. Apply the lemma of the above exercise to $\varrho(\zeta)/(\zeta - 1)$ and show that the BDF-formulas are stable for $k = 1, 2, 3, 4$.
4. Give a different proof of Theorem 3.4 by applying the Schur-Cohn criterion to the polynomial

$$f(z) = z^k \varrho\left(\frac{1}{z}\right) = \sum_{j=1}^k \frac{1}{j} (1-z)^j. \quad (3.25)$$

Schur-Cohn criterion (see e.g., Marden (1966), Chapter X). For a given polynomial with real coefficients

$$f(z) = a_0 + a_1 z + \dots + a_k z^k$$

we consider the coefficients $a_i^{(j)}$ where

$$\begin{aligned} a_i^{(0)} &= a_i & i &= 0, 1, \dots, k \\ a_i^{(j+1)} &= a_0^{(j)} a_i^{(j)} - a_{k-j}^{(j)} a_{k-j-i}^{(j)} & i &= 0, 1, \dots, k-j-1 \end{aligned} \quad (3.26)$$

and also the products

$$P_1 = a_0^{(1)}, \quad P_{j+1} = P_j a_0^{(j+1)} \quad \text{for } j = 1, \dots, k-1. \quad (3.27)$$

We further denote by n the number of negative elements among the values P_1, \dots, P_k and by p the number of positive elements. Then $f(z)$ has at least n zeros inside the unit disk and at least p zeros outside it.

- a) Prove the following formulas for the coefficients of (3.25):

$$\begin{aligned} a_0 &= \sum_{i=1}^k \frac{1}{i}, & a_1 &= -k, & a_2 &= \frac{k(k-1)}{4}, \\ a_{k-2} &= (-1)^k \frac{k(k-1)}{2(k-2)}, & a_{k-1} &= (-1)^{k-1} \frac{k}{k-1}, & a_k &= (-1)^k \frac{1}{k}. \end{aligned} \quad (3.28)$$

- b) Verify that the coefficients $a_0^{(j)}$ of (3.26) have the sign structure of Table 3.1. For $k < 13$ these tedious calculations can be performed on a computer. The verification of $a_0^{(1)} > 0$ and $a_0^{(2)} > 0$ is easy for all $k > 2$. In order to verify $a_0^{(3)} = (a_0^{(2)})^2 - (a_{k-2}^{(2)})^2 < 0$ for $k \geq 13$ consider the expression

$$\begin{aligned} a_0^{(2)} - (-1)^k a_{k-2}^{(2)} &= a_0^{(1)} (a_0^2 - a_k^2 - a_0 |a_{k-2}| + a_2 |a_k|) \\ &\quad - |a_{k-1}^{(1)}| \cdot (a_0 + |a_k|) (|a_{k-1}| + a_1) \end{aligned} \quad (3.29)$$

Table 3.1. Signs of $a_0^{(j)}$.

k	2	3	4	5	6	7	8	9	10	11	12	13	> 13
$j=1$	+	+	+	+	+	+	+	+	+	+	+	+	+
$j=2$	0	+	+	+	+	+	+	+	+	+	+	+	+
$j=3$		0	+	+	+	+	+	+	+	+	+	-	-
$j=4$			0	+	+	+	-	-	-	-	-		
$j=5$				0	+	-							

which can be written in the form $(a_0 + |a_k|)\varphi(k)$ with

$$\begin{aligned}
 \varphi(k) &= (a_0 - |a_k|)(a_0^2 - a_k^2 - a_0|a_{k-2}| + a_2|a_k|) - |a_{k-1}^{(1)}|(a_1 + |a_{k-1}|) \\
 &= a_0^3 - a_0^2\left(\frac{k}{2} + \frac{1}{2} + \frac{1}{k-2} + \frac{1}{k}\right) \\
 &\quad + a_0\left(\frac{5k}{4} + \frac{1}{4} + \frac{1}{2k-4} - \frac{1}{k-1} - \frac{1}{(k-1)^2} - \frac{1}{k^2}\right) \\
 &\quad - \left(k - \frac{3}{4} - \frac{1}{k-1} - \frac{1}{4k} - \frac{1}{k^3}\right).
 \end{aligned}$$

Show that $\varphi(13) < 0$ and that φ is monotonically decreasing for $k \geq 13$ (observe that $a_0 = a_0(k)$ actually depends on k and that $a_0(k+1) = a_0(k) + 1/(k+1)$). Finally, deduce from the negativeness of (3.29) that $a_0^{(3)} < 0$ for $k \geq 13$.

- c) Use Table 3.1 and the Schur-Cohn criterion for the verification of Theorem 3.4.
5. (Multistep methods of maximal order). Verify the following statements:
 - a) there is no k -step method of order $2k+1$,
 - b) there is a unique (implicit) k -step method of order $2k$,
 - c) there is a unique explicit k -step method of order $2k-1$.
6. Prove that symmetric multistep methods are always of even order. More precisely, if a symmetric multistep method is of order $2s-1$ then it is also of order $2s$.
7. Show that all stable 4-step methods of order 6 are given by

$$\begin{aligned}
 \varrho(\zeta) &= (\zeta^2 - 1)(\zeta^2 + 2\mu\zeta + 1), \quad |\mu| < 1, \\
 \sigma(\zeta) &= \frac{1}{45}(14 - \mu)(\zeta^4 + 1) + \frac{1}{45}(64 + 34\mu)\zeta(\zeta^2 + 1) + \frac{1}{15}(8 + 38\mu)\zeta^2.
 \end{aligned}$$

Compute the error constant and observe that it cannot become arbitrarily small.

Result. $C = -(16 - 5\mu)/(7560(1 + \mu))$.

8. Prove the following bounds for the error constant:

a) For stable methods of order $k + 2$

$$C \leq -2^{-1-k} \mu_{k+1}.$$

b) For stable methods of order $k + 1$ with odd k we have

$$C \leq -2^{-k} \mu_k.$$

c) For stable explicit methods of order k we have ($\mu_j = 0$ for even j)

$$C \geq 2^{1-k} \left(\frac{1}{2} - \sum_{j=1}^{k-1} \mu_j \right).$$

Show that all these bounds are optimal.

Hint. Compare the formulas (3.18) and (3.21) and use the relation $\sigma(1) = 2^{k-1} a_{k-1}$ of Lemma 3.6.

9. The coefficients μ_j of formula (3.19) satisfy the recurrence relation

$$\mu_{2j+1} + \frac{1}{3} \mu_{2j-1} + \dots + \frac{1}{2j+1} \mu_1 = \frac{1}{4j+6}. \quad (3.30)$$

The first of these coefficients are given by

$$\mu_1 = \frac{1}{6}, \quad \mu_3 = \frac{2}{45}, \quad \mu_5 = \frac{22}{945}, \quad \mu_7 = \frac{214}{14175}.$$

10. Another proof of Lemma 3.8: multiplying (3.30) by $2j + 3$ and subtracting from it the same formula with j replaced by $j - 1$ yields

$$(2j+3)\mu_{2j+1} + \sum_{i=0}^{j-1} \mu_{2i+1} \left(\frac{2j+3}{2j-2i+1} - \frac{2j+1}{2j-2i-1} \right) = 0.$$

Show that the expression in brackets is negative and deduce the result of Lemma 3.8 by a simple induction argument.

III.4 Convergence of Multistep Methods

... , ist das Adams'sche Verfahren jedem andern bedeutend überlegen. Wenn es gleichwohl nicht genügend allgemein angewandt wird und, besonders in Deutschland, gegenüber den von Runge, Heun und Kutta entwickelten Methoden zurücktritt, so mag dies daran liegen, dass bisher eine brauchbare Untersuchung der Genauigkeit der Adams'schen Integration gefehlt hat. Diese Lücke soll hier ausgefüllt werden, ...

(R. v. Mises 1930)

The convergence of Adams methods was investigated in the influential article of von Mises (1930), which was followed by an avalanche of papers improving the error bounds and applying the ideas to other special multistep methods, e.g., Tollmien (1938), Fricke (1949), Weissinger (1950), Vietoris (1953). A general convergence proof for the method (2.1), however, was first given by Dahlquist (1956), who gave necessary and sufficient conditions for convergence. Great elegance was introduced in the proofs by the ideas of Butcher (1966), where multistep formulas are written as one-step formulas in a higher dimensional space. Furthermore, the resulting presentation can easily be extended to a more general class of integration methods (see Section III.8).

We cannot expect reasonable convergence of numerical methods, if the differential equation problem

$$y' = f(x, y), \quad y(x_0) = y_0 \quad (4.1)$$

does not possess a unique solution. We therefore make the following assumptions, which were seen in Sections I.7 and I.9 to be natural for our purpose:

$$f \text{ is continuous on } D = \{(x, y) ; x \in [x_0, \hat{x}], \|y(x) - y\| \leq b\} \quad (4.2a)$$

where $y(x)$ denotes the exact solution of (4.1) and b is some positive number. We further assume that f satisfies a Lipschitz condition, i.e.,

$$\|f(x, y) - f(x, z)\| \leq L\|y - z\| \quad \text{for } (x, y), (x, z) \in D. \quad (4.2b)$$

If we apply the multistep method (2.1) with step size h to the problem (4.1) we obtain a sequence $\{y_i\}$. For given x and h such that $(x - x_0)/h = n$ is an integer, we introduce the following notation for the numerical solution:

$$y_h(x) = y_n \quad \text{if } x - x_0 = nh. \quad (4.3)$$

Definition 4.1 (Convergence). i) The linear multistep method (2.1) is called *convergent*, if for all initial value problems (4.1) satisfying (4.2),

$$y(x) - y_h(x) \rightarrow 0 \quad \text{for } h \rightarrow 0, x \in [x_0, \hat{x}]$$

whenever the starting values satisfy

$$y(x_0 + ih) - y_h(x_0 + ih) \rightarrow 0 \quad \text{for } h \rightarrow 0, i = 0, 1, \dots, k-1.$$

ii) Method (2.1) is *convergent of order p* , if to any problem (4.1) with f sufficiently differentiable, there exists a positive h_0 such that

$$\|y(x) - y_h(x)\| \leq Ch^p \quad \text{for } h \leq h_0$$

whenever the starting values satisfy

$$\|y(x_0 + ih) - y_h(x_0 + ih)\| \leq C_0 h^p \quad \text{for } h \leq h_0, i = 0, 1, \dots, k-1.$$

In this definition we clearly assume that a solution of (4.1) exists on $[x_0, \hat{x}]$.

The aim of this section is to prove that stability together with consistency are necessary and sufficient for the convergence of a multistep method. This is expressed in the famous slogan

$$\text{convergence} = \text{stability} + \text{consistency}$$

(compare also Lax & Richtmyer 1956). We begin with the study of necessary conditions for convergence.

Theorem 4.2. *If the multistep method (2.1) is convergent, then it is necessarily*

- i) *stable and*
- ii) *consistent (i.e. of order 1: $\varrho(1) = 0$, $\varrho'(1) = \sigma(1)$).*

Proof. Application of the multistep method (2.1) to the differential equation $y' = 0$, $y(0) = 0$ yields the difference equation (3.6). Suppose, by contradiction, that $\varrho(\zeta)$ has a root ζ_1 with $|\zeta_1| > 1$, or a root ζ_2 on the unit circle whose multiplicity exceeds 1. ζ_1^n and $n\zeta_2^n$ are then divergent solutions of (3.6). Multiplying by \sqrt{h} we achieve that the starting values converge to $y_0 = 0$ for $h \rightarrow 0$. Since $y_h(x) = \sqrt{h}\zeta_1^{x/h}$ and $y_h(x) = (x/\sqrt{h})\zeta_2^{x/h}$ remain divergent for every fixed x , we have a contradiction to the assumption of convergence. The method (2.1) must therefore be stable.

We next consider the initial value problem $y' = 0$, $y(0) = 1$ with exact solution $y(x) = 1$. The corresponding difference equation is again that of (3.6), which, in the new notation, can be written as

$$\alpha_k y_h(x + kh) + \alpha_{k-1} y_h(x + (k-1)h) + \dots + \alpha_0 y_h(x) = 0.$$

Letting $h \rightarrow 0$, convergence immediately implies that $\varrho(1) = 0$.

Finally we apply method (2.1) to the problem $y' = 1$, $y(0) = 0$. The exact solution is $y(x) = x$. Since we already know that $\varrho(1) = 0$, it is easy to verify that a particular numerical solution is given by $y_n = nhK$ or $y_h(x) = xK$ where $K = \sigma(1)/\varrho'(1)$. By convergence, $K = 1$ is necessary. \square

Although the statement of Theorem 4.2 was derived from a consideration of almost trivial differential equations, it is remarkable that conditions (i) and (ii) turn out to be not only necessary but also sufficient for convergence.

Formulation as One-Step Method

We are now at the point where it is useful to rewrite a multistep method as a one-step method in a higher dimensional space (see Butcher 1966, Skeel 1976). For this let $\psi = \psi(x_i, y_i, \dots, y_{i+k-1}, h)$ be defined implicitly by

$$\psi = \sum_{j=0}^{k-1} \beta'_j f(x_i + jh, y_{i+j}) + \beta'_k f\left(x_i + kh, h\psi - \sum_{j=0}^{k-1} \alpha'_j y_{i+j}\right) \quad (4.4)$$

where $\alpha'_j = \alpha_j/\alpha_k$ and $\beta'_j = \beta_j/\alpha_k$. Multistep formula (2.1) can then be written as

$$y_{i+k} = - \sum_{j=0}^{k-1} \alpha'_j y_{i+j} + h\psi. \quad (4.5)$$

Introducing the $m \cdot k$ -dimensional vectors (m is the dimension of the differential equation)

$$Y_i = (y_{i+k-1}, y_{i+k-2}, \dots, y_i)^T, \quad i \geq 0 \quad (4.6)$$

and

$$A = \begin{pmatrix} -\alpha'_{k-1} & -\alpha'_{k-2} & \cdots & \cdot & -\alpha'_0 \\ 1 & 0 & \cdots & \cdot & 0 \\ & 1 & & \cdot & 0 \\ & & \ddots & \vdots & \vdots \\ & & & 1 & 0 \end{pmatrix}, \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad (4.7)$$

the multistep method (4.5) can be written — after adding some trivial identities — in compact form as

$$Y_{i+1} = (A \otimes I)Y_i + h\Phi(x_i, Y_i, h), \quad i \geq 0 \quad (4.8)$$

with

$$\Phi(x_i, Y_i, h) = (e_1 \otimes I)\psi(x_i, Y_i, h). \quad (4.8a)$$

Here, $A \otimes I$ denotes the Kronecker tensor product, i.e. the $m \cdot k$ -dimensional block matrix with (m, m) -blocks $a_{ij}I$. Readers unfamiliar with the notation and properties of this product may assume for simplicity that (4.1) is a scalar equation ($m=1$) and $A \otimes I = A$.

The following lemmas express the concepts of order and stability in this new notation.

Lemma 4.3. Let $y(x)$ be the exact solution of (4.1). For $i = 0, 1, 2, \dots$ we define the vector \hat{Y}_{i+1} as the numerical solution of one step

$$\hat{Y}_{i+1} = (A \otimes I)Y(x_i) + h\Phi(x_i, Y(x_i), h)$$

with correct starting values

$$Y(x_i) = (y(x_{i+k-1}), y(x_{i+k-2}), \dots, y(x_i))^T.$$

i) If the multistep method (2.1) is of order 1 and if f satisfies (4.2), then an $h_0 > 0$ exists such that for $h \leq h_0$,

$$\|Y(x_{i+1}) - \hat{Y}_{i+1}\| \leq h\omega(h), \quad 0 \leq i \leq \hat{x}/h - k$$

where $\omega(h) \rightarrow 0$ for $h \rightarrow 0$.

ii) If the multistep method (2.1) is of order p and if f is sufficiently differentiable then a constant M exists such that for h small enough,

$$\|Y(x_{i+1}) - \hat{Y}_{i+1}\| \leq Mh^{p+1}, \quad 0 \leq i \leq \hat{x}/h - k.$$

Proof. The first component of $Y(x_{i+1}) - \hat{Y}_{i+1}$ is the local error as given by Definition 2.1. Since the remaining components all vanish, Exercise 5 of Section III.2 and Definition 2.3 yield the result. \square

Lemma 4.4. Suppose that the multistep method (2.1) is stable. Then there exists a vector norm (on \mathbb{R}^{mk}) such that the matrix A of (4.7) satisfies

$$\|A \otimes I\| \leq 1$$

in the subordinate matrix norm.

Proof. If λ is a root of $\varrho(\zeta)$, then the vector $(\lambda^{k-1}, \lambda^{k-2}, \dots, 1)$ is an eigenvector of the matrix A with eigenvalue λ . Therefore the eigenvalues of A (which are the roots of $\varrho(\zeta)$) satisfy the root condition by Definition 3.2. A transformation to Jordan canonical form therefore yields (see Section I.12)

$$T^{-1}AT = J = \text{diag} \left\{ \lambda_1, \dots, \lambda_l, \begin{pmatrix} \lambda_{l+1} & \varepsilon_{l+1} & & \\ & \ddots & \varepsilon_{k-1} & \\ & & \lambda_k & \end{pmatrix} \right\} \quad (4.9)$$

where $\lambda_1, \dots, \lambda_l$ are the eigenvalues of modulus 1, which must be simple, each ε_j is either 0 or 1. We further find by a suitable multiplication of the columns of T that $|\varepsilon_j| < 1 - |\lambda_j|$ for $j = l+1, \dots, k-1$. Because of (9.11') of Chapter I we then have $\|J \otimes I\|_\infty \leq 1$. Using the transformation T of (4.9) we define the norm

$$\|x\| := \|(T^{-1} \otimes I)x\|_\infty.$$

This yields

$$\begin{aligned}\|(A \otimes I)x\| &= \|(T^{-1} \otimes I)(A \otimes I)x\|_{\infty} = \|(J \otimes I)(T^{-1} \otimes I)x\|_{\infty} \\ &\leq \|(T^{-1} \otimes I)x\|_{\infty} = \|x\|\end{aligned}$$

and hence also $\|A \otimes I\| \leq 1$. \square

Proof of Convergence

The convergence theorem for multistep methods can now be established.

Theorem 4.5. *If the multistep method (2.1) is stable and of order 1 then it is convergent. If method (2.1) is stable and of order p then it is convergent of order p .*

Proof. As in the convergence theorem for one-step methods (Section II.3) we may assume without loss of generality that $f(x, y)$ is defined for all $y \in \mathbb{R}^m$, $x \in [x_0, \hat{x}]$ and satisfies there a (global) Lipschitz condition. This implies that for sufficiently small h the functions $\psi(x_i, Y_i, h)$ and $\Phi(x_i, Y_i, h)$ satisfy a Lipschitz condition with respect to the second argument (with Lipschitz constant L^*). For the function G , defined by formula (4.8), which maps the vector Y_i onto Y_{i+1} we thus obtain from Lemma 4.4

$$\|G(Y_i) - G(Z_i)\| \leq (1 + hL^*)\|Y_i - Z_i\|. \quad (4.10)$$

The rest of the proof now proceeds in the same way as for one-step methods and is illustrated in Fig. 4.1.

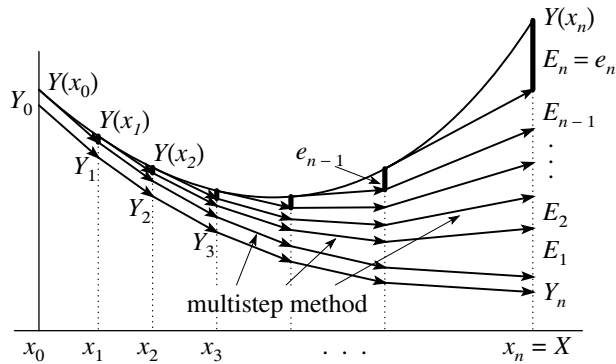


Fig. 4.1. Lady Windermere's Fan for multistep methods

The arrows in Fig. 4.1 indicate the application of G . From Lemma 4.3 we know that $\|Y(x_{i+1}) - G(Y(x_i))\| \leq h\omega(h)$. This together with (4.10) shows that

the local error $Y(x_{i+1}) - G(Y(x_i))$ at stage $i+1$ causes an error at stage n , which is at most $h\omega(h)(1+hL^*)^{n-i+1}$. Thus we have

$$\begin{aligned} \|Y(x_n) - Y_n\| &\leq \|Y(x_0) - Y_0\|(1+hL^*)^n \\ &\quad + h\omega(h)\left((1+hL^*)^{n-1} + (1+hL^*)^{n-2} + \dots + 1\right) \\ &\leq \|Y(x_0) - Y_0\| \exp(nhL^*) + \frac{\omega(h)}{L^*}(\exp(nhL^*) - 1). \end{aligned} \quad (4.11)$$

Convergence of method (2.1) is now an immediate consequence of formula (4.11). If the multistep method is of order p , the same proof with $\omega(h)$ replaced by Mh^p yields convergence of order p . \square

Exercises

1. Consider the function (for $x \geq 0$)

$$f(x, y) = \begin{cases} 2x & \text{for } y \leq 0, \\ 2x - \frac{4y}{x} & \text{for } 0 < y < x^2, \\ -2x & \text{for } y \geq x^2. \end{cases}$$

- Show that $y(x) = x^2/3$ is the unique solution of $y' = f(x, y)$, $y(0) = 0$, although f does not satisfy a Lipschitz condition near the origin.
 - Apply the mid-point rule (1.13') with starting values $y_0 = 0$, $y_1 = -h^2$ to the above problem and verify that the numerical solution at $x = nh$ is given by $y_h(x) = (-1)^n x^2$ (Taubert 1976, see also Grigorieff 1977).
2. Another motivation for the meaning of the error constant: suppose that 1 is the only eigenvalue of A in (4.7) of modulus one. Show that $(1, 1, \dots, 1)^T$ is the right eigenvector and $(1, 1 + \alpha'_{k-1}, 1 + \alpha'_{k-1} + \alpha'_{k-2}, \dots)$ is the left eigenvector to this eigenvalue. The *global* contribution of the *local* error after many steps is then given by

$$A^\infty \begin{pmatrix} C_{p+1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} = C \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \quad (4.12)$$

Multiply this equation from the left by the left eigenvector to show with (2.6) that C is the error constant defined in (2.13).

Remark. For multistep methods with several eigenvalues of modulus 1, formula (4.12) remains valid if A^∞ is replaced by E (see Section III.8).

III.5 Variable Step Size Multistep Methods

Des war a harter Brockn, des . . . (Tyrolean dialect)

It is clear from the considerations of Section II.4 that an efficient integrator must be able to change the step size. However, changing the step size with multistep methods is difficult since the formulas of the preceding sections require the numerical approximations at equidistant points. There are in principle two possibilities for overcoming this difficulty:

- i) use polynomial interpolation to reproduce the starting values at the new (equidistant) grid;
- ii) construct methods which are adjusted to variable grid points.

This section is devoted to the second approach. We investigate consistency, stability and convergence. The actual implementation (order and step size strategies) will be considered in Section III.7.

Variable Step Size Adams Methods

F. Ceschino (1961) was apparently the first person to propose a “smooth” transition from a step size h to a new step size ωh . C.V.D. Forrington (1961) and later on F.T. Krogh (1969) extended his ideas: we consider an arbitrary grid (x_n) and denote the step sizes by $h_n = x_{n+1} - x_n$. We assume that approximations y_j to $y(x_j)$ are known for $j = n - k + 1, \dots, n$ and we put $f_j = f(x_j, y_j)$. In the same way as in Section III.1 we denote by $p(t)$ the polynomial which interpolates the values (x_j, f_j) for $j = n - k + 1, \dots, n$. Using Newton’s interpolation formula we have

$$p(t) = \sum_{j=0}^{k-1} \prod_{i=0}^{j-1} (t - x_{n-i}) \delta^j f[x_n, x_{n-1}, \dots, x_{n-j}] \quad (5.1)$$

where the divided differences $\delta^j f[x_n, \dots, x_{n-j}]$ are defined recursively by

$$\begin{aligned} \delta^0 f[x_n] &= f_n \\ \delta^j f[x_n, \dots, x_{n-j}] &= \frac{\delta^{j-1} f[x_n, \dots, x_{n-j+1}] - \delta^{j-1} f[x_{n-1}, \dots, x_{n-j}]}{x_n - x_{n-j}}. \end{aligned} \quad (5.2)$$

For actual computations (see Krogh 1969) it is practical to rewrite (5.1) as

$$p(t) = \sum_{j=0}^{k-1} \prod_{i=0}^{j-1} \frac{t - x_{n-i}}{x_{n+1} - x_{n-i}} \cdot \Phi_j^*(n) \quad (5.1')$$

where

$$\Phi_j^*(n) = \prod_{i=0}^{j-1} (x_{n+1} - x_{n-i}) \cdot \delta^j f[x_n, \dots, x_{n-j}]. \quad (5.3)$$

We now define the approximation to $y(x_{n+1})$ by

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} p(t) dt. \quad (5.4)$$

Inserting formula (5.1') into (5.4) we obtain

$$y_{n+1} = y_n + h_n \sum_{j=0}^{k-1} g_j(n) \Phi_j^*(n) \quad (5.5)$$

with

$$g_j(n) = \frac{1}{h_n} \int_{x_n}^{x_{n+1}} \prod_{i=0}^{j-1} \frac{t - x_{n-i}}{x_{n+1} - x_{n-i}} dt. \quad (5.6)$$

Formula (5.5) is the extension of the explicit Adams method (1.5) to variable step sizes. Observe that for constant step sizes the above expressions reduce to (Exercise 1)

$$g_j(n) = \gamma_j, \quad \Phi_j^*(n) = \nabla^j f_n.$$

The variable step size *implicit* Adams methods can be deduced similarly. In analogy to Section III.1 we let $p^*(t)$ be the polynomial of degree k that interpolates (x_j, f_j) for $j = n - k + 1, \dots, n, n + 1$ (the value $f_{n+1} = f(x_{n+1}, y_{n+1})$ contains the unknown solution y_{n+1}). Again, using Newton's interpolation formula we obtain

$$p^*(t) = p(t) + \prod_{i=0}^{k-1} (t - x_{n-i}) \cdot \delta^k f[x_{n+1}, x_n, \dots, x_{n-k+1}].$$

The numerical solution, defined by

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} p^*(t) dt,$$

is now given by

$$y_{n+1} = p_{n+1} + h_n g_k(n) \Phi_k^*(n+1), \quad (5.7)$$

where p_{n+1} is the numerical approximation obtained by the explicit Adams method

$$p_{n+1} = y_n + h_n \sum_{j=0}^{k-1} g_j(n) \Phi_j^*(n)$$

and where

$$\Phi_k(n+1) = \prod_{i=0}^{k-1} (x_{n+1} - x_{n-i}) \cdot \delta^k f[x_{n+1}, x_n, \dots, x_{n-k+1}]. \quad (5.8)$$

Recurrence Relations for $g_j(n)$, $\Phi_j(n)$ and $\Phi_j^*(n)$

The cost of computing integration coefficients is the biggest disadvantage to permitting arbitrary variations in the step size.

(F.T. Krogh 1973)

The values $\Phi_j^*(n)$ ($j = 0, \dots, k-1$) and $\Phi_k(n+1)$ can be computed efficiently with the recurrence relations

$$\begin{aligned} \Phi_0(n) &= \Phi_0^*(n) = f_n \\ \Phi_{j+1}(n) &= \Phi_j(n) - \Phi_j^*(n-1) \\ \Phi_j^*(n) &= \beta_j(n) \Phi_j(n), \end{aligned} \quad (5.9)$$

which are an immediate consequence of Definitions (5.3) and (5.8). The coefficients

$$\beta_j(n) = \prod_{i=0}^{j-1} \frac{x_{n+1} - x_{n-i}}{x_n - x_{n-i-1}}$$

can be calculated by

$$\beta_0(n) = 1, \quad \beta_j(n) = \beta_{j-1}(n) \frac{x_{n+1} - x_{n-j+1}}{x_n - x_{n-j}}.$$

The calculation of the coefficients $g_j(n)$ is trickier (F.T. Krogh 1974). We introduce the q -fold integral

$$c_{jq}(x) = \frac{(q-1)!}{h_n^q} \int_{x_n}^x \int_{x_n}^{\xi_{q-1}} \dots \int_{x_n}^{\xi_1} \prod_{i=0}^{j-1} \frac{\xi_0 - x_{n-i}}{x_{n+1} - x_{n-i}} d\xi_0 \dots d\xi_{q-1} \quad (5.10)$$

and observe that

$$g_j(n) = c_{j1}(x_{n+1}).$$

Lemma 5.1. *We have*

$$\begin{aligned} c_{0q}(x_{n+1}) &= \frac{1}{q}, & c_{1q}(x_{n+1}) &= \frac{1}{q(q+1)}, \\ c_{jq}(x_{n+1}) &= c_{j-1,q}(x_{n+1}) - c_{j-1,q+1}(x_{n+1}) \frac{h_n}{x_{n+1} - x_{n-j+1}}. \end{aligned}$$

Proof. The first two relations follow immediately from (5.10). In order to prove the recurrence relation we denote by $d(x)$ the difference

$$d(x) = c_{jq}(x) - c_{j-1,q}(x) \frac{x - x_{n-j+1}}{x_{n+1} - x_{n-j+1}} + c_{j-1,q+1}(x) \frac{h_n}{x_{n+1} - x_{n-j+1}}.$$

Clearly, $d^{(i)}(x_n) = 0$ for $i = 0, 1, \dots, q-1$. Moreover, the q -th derivative of $d(x)$ vanishes, since by the Leibniz rule

$$\begin{aligned} \frac{d^q}{dx^q} \left(c_{j-1,q}(x) \cdot \frac{x - x_{n-j+1}}{x_{n+1} - x_{n-j+1}} \right) \\ = c_{j-1,q}^{(q)}(x) \frac{x - x_{n-j+1}}{x_{n+1} - x_{n-j+1}} + q c_{j-1,q}^{(q-1)}(x) \frac{1}{x_{n+1} - x_{n-j+1}} \\ = c_{j,q}^{(q)}(x) + c_{j-1,q+1}^{(q)}(x) \frac{h_n}{x_{n+1} - x_{n-j+1}}. \end{aligned}$$

Therefore we have $d(x) \equiv 0$ and the statement follows by putting $x = x_{n+1}$. \square

Using the above recurrence relation one can successively compute $c_{2q}(x_{n+1})$ for $q = 1, \dots, k-1$; $c_{3q}(x_{n+1})$ for $q = 1, \dots, k-2$; \dots ; $c_{kq}(x_{n+1})$ for $q = 1$. This procedure yields in an efficient way the coefficients $g_j(n) = c_{j1}(x_{n+1})$ of the Adams methods.

Variable Step Size BDF

The BDF-formulas (1.22) can also be extended in a natural way to variable step size. Denote by $q(t)$ the polynomial of degree k that interpolates (x_i, y_i) for $i = n+1, n, \dots, n-k+1$. It can be expressed, using divided differences, by

$$q(t) = \sum_{j=0}^k \prod_{i=0}^{j-1} (t - x_{n+1-i}) \cdot \delta^j y[x_{n+1}, x_n, \dots, x_{n-j+1}]. \quad (5.11)$$

The requirement

$$q'(x_{n+1}) = f(x_{n+1}, y_{n+1})$$

immediately leads to the variable step size BDF-formulas

$$\sum_{j=1}^k h_n \prod_{i=1}^{j-1} (x_{n+1} - x_{n+1-i}) \cdot \delta^j y[x_{n+1}, \dots, x_{n-j+1}] = h_n f(x_{n+1}, y_{n+1}). \quad (5.12)$$

The computation of the coefficients is much easier here than for the Adams methods.

General Variable Step Size Methods and Their Orders

For theoretical investigations it is convenient to write the methods in a form where the y_j and f_j values appear linearly. For example, the implicit Adams method (5.7) becomes ($k = 2$)

$$y_{n+1} = y_n + \frac{h_n}{6(1+\omega_n)} \left((3+2\omega_n)f_{n+1} + (3+\omega_n)(1+\omega_n)f_n - \omega_n^2 f_{n-1} \right), \quad (5.13)$$

where we have introduced the notation $\omega_n = h_n/h_{n-1}$ for the step size ratio. Or, the 2-step BDF-formula (5.12) can be written as

$$y_{n+1} - \frac{(1+\omega_n)^2}{1+2\omega_n} y_n + \frac{\omega_n^2}{1+2\omega_n} y_{n-1} = h_n \frac{1+\omega_n}{1+2\omega_n} f_{n+1}. \quad (5.14)$$

In order to give a unified theory for all these variable step size multistep methods we consider formulas of the form

$$y_{n+k} + \sum_{j=0}^{k-1} \alpha_{jn} y_{n+j} = h_{n+k-1} \sum_{j=0}^k \beta_{jn} f_{n+j}. \quad (5.15)$$

The coefficients α_{jn} and β_{jn} actually depend on the ratios $\omega_i = h_i/h_{i-1}$ for $i = n+1, \dots, n+k-1$. In analogy to the constant step size case we give

Definition 5.2. Method (5.15) is *consistent of order p* , if

$$q(x_{n+k}) + \sum_{j=0}^{k-1} \alpha_{jn} q(x_{n+j}) = h_{n+k-1} \sum_{j=0}^k \beta_{jn} q'(x_{n+j})$$

holds for all polynomials $q(x)$ of degree $\leq p$ and for all grids (x_j) .

By definition, the explicit Adams method (5.5) is of order k , the implicit Adams method (5.7) is of order $k+1$, and the BDF-formula (5.12) is of order k .

The notion of consistency certainly has to be related to the local error. Indeed, if the method is of order p , if the ratios h_j/h_n are bounded for $j = n+1, \dots, n+k-1$ and if the coefficients satisfy

$$\alpha_{jn}, \beta_{jn} \text{ are bounded,} \quad (5.16)$$

then a Taylor expansion argument implies that

$$y(x_{n+k}) + \sum_{j=0}^{k-1} \alpha_{jn} y(x_{n+j}) - h_{n+k-1} \sum_{j=0}^k \beta_{jn} y'(x_{n+j}) = \mathcal{O}(h_n^{p+1}) \quad (5.17)$$

for sufficiently smooth $y(x)$. Interpreting $y(x)$ as the solution of the differential equation, a trivial extension of Lemma 2.2 to variable step sizes shows that the local error at x_{n+k} (cf. Definition 2.1) is also $\mathcal{O}(h_n^{p+1})$.

This motivates the investigation of condition (5.16). The methods (5.13) and (5.14) are seen to satisfy (5.16) whenever the step size ratio h_n/h_{n-1} is bounded from above. In general we have

Lemma 5.3. *For the explicit and implicit Adams methods as well as for the BDF-formulas the coefficients α_{jn} and β_{jn} are bounded whenever for some Ω*

$$h_n/h_{n-1} \leq \Omega.$$

Proof. We prove the statement for the explicit Adams methods only. The proof for the other methods is similar and thus omitted. We see from formula (5.5) that the coefficients α_{jn} do not depend on n and hence are bounded. The β_{jn} are composed of products of $g_j(n)$ with the coefficients of $\Phi_j^*(n)$, when written as a linear combination of f_n, \dots, f_{n-j} . From formula (5.6) we see that $|g_j(n)| \leq 1$. It follows from $(x_{n+1} - x_{n-j+1}) \leq \max(1, \Omega^j)(x_n - x_{n-j})$ and from an induction argument that the coefficients of $\Phi_j^*(n)$ are also bounded. Hence the β_{jn} are bounded, which proves the lemma. \square

The condition $h_n/h_{n-1} \leq \Omega$ is a reasonable assumption which can easily be satisfied by a code.

Stability

So geht das einfach . . . (R.D. Grigorieff, Halle 1983)

The study of stability for variable step size methods was begun in the articles of Gear & Tu (1974) and Gear & Watanabe (1974). Further investigations are due to Grigorieff (1983) and Crouzeix & Lisbona (1984).

We have seen in Section III.3 that for equidistant grids stability is equivalent to the boundedness of the numerical solution, when applied to the scalar differential equation $y' = 0$. Let us do the same here for the general case. Method (5.15), applied to $y' = 0$, gives the difference equation with variable coefficients

$$y_{n+k} + \sum_{j=0}^{k-1} \alpha_{jn} y_{n+j} = 0.$$

If we introduce the vector $Y_n = (y_{n+k-1}, \dots, y_n)^T$, this difference equation is equivalent to

$$Y_{n+1} = A_n Y_n$$

with

$$A_n = \begin{pmatrix} -\alpha_{k-1,n} & \cdots & \cdots & -\alpha_{1,n} & -\alpha_{0,n} \\ 1 & 0 & \cdots & 0 & 0 \\ & \ddots & \ddots & \vdots & \vdots \\ & & 1 & 0 & 0 \\ & & & 1 & 0 \end{pmatrix}, \quad (5.18)$$

the companion matrix.

Definition 5.4. Method (5.15) is called *stable*, if

$$\|A_{n+l}A_{n+l-1} \cdots A_{n+1}A_n\| \leq M \quad (5.19)$$

for all n and $l \geq 0$.

Observe that in general A_n depends on the step ratios $\omega_{n+1}, \dots, \omega_{n+k-1}$. Therefore, condition (5.19) will usually lead to a restriction on these values. For the Adams methods (5.5) and (5.7) the coefficients α_{jn} do not depend on n and hence are stable for any step size sequence.

In the following three theorems we present stability results for general variable step size methods. The first one, taken from Crouzeix & Lisbona (1984), is a sort of perturbation result: the variable step size method is considered as a perturbation of a strongly stable fixed step size method.

Theorem 5.5. *Let the method (5.15) satisfy the following properties:*

- a) *it is of order $p \geq 0$, i.e., $1 + \sum_{j=0}^{k-1} \alpha_{jn} = 0$;*
- b) *the coefficients $\alpha_{jn} = \alpha_j(\omega_{n+1}, \dots, \omega_{n+k-1})$ are continuous in a neighbourhood of $(1, \dots, 1)$;*
- c) *the underlying constant step size formula is strongly stable, i.e., all roots of*

$$\zeta^k + \sum_{j=0}^{k-1} \alpha_j(1, \dots, 1)\zeta^j = 0$$

lie in the open unit disc $|\zeta| < 1$, with the exception of $\zeta_1 = 1$.

Then there exist real numbers ω, Ω ($\omega < 1 < \Omega$) such that the method is stable if

$$\omega \leq h_n/h_{n-1} \leq \Omega \quad \text{for all } n. \quad (5.20)$$

Proof. Let A be the companion matrix of the constant step size formula. As in the proof of Lemma 4.4 we transform A to Jordan canonical form and obtain

$$T^{-1}AT = \begin{pmatrix} \hat{A} & 0 \\ & 1 \end{pmatrix}$$

where, by assumption (c), $\|\hat{A}\|_1 < 1$. Observe that the last column of T , the eigenvector of A corresponding to 1, is given by $t_k = (1, \dots, 1)^T$. Assumption (a) implies that this vector t_k is also an eigenvector for each A_n . Therefore we have

$$T^{-1}A_nT = \begin{pmatrix} \hat{A}_n & \begin{smallmatrix} 0 \\ \vdots \\ 0 \end{smallmatrix} \\ \hat{A}_n & 1 \end{pmatrix}$$

and, by continuity, $\|\hat{A}_n\|_1 \leq 1$, if $\omega_{n+1}, \dots, \omega_{n+k-1}$ are sufficiently close to 1. Stability now follows from the fact that

$$\|T^{-1}A_nT\|_1 = \max(\|\hat{A}_n\|_1, 1) = 1,$$

which implies that

$$\|A_{n+l} \dots A_{n+1} A_n\| \leq \|T\| \cdot \|T^{-1}\|. \quad \square$$

The next result (Grigorieff 1983) is based on a reduction of the dimension of the matrices A_n by one. The idea is to use the transformation

$$T = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ & 1 & 1 & \dots & 1 \\ & & 1 & \dots & 1 \\ & & & \ddots & \vdots \\ 0 & & & & 1 \end{pmatrix}, \quad T^{-1} = \begin{pmatrix} 1 & -1 & & 0 & \\ & 1 & -1 & & \\ & & 1 & \ddots & \\ & & & 1 & -1 \\ 0 & & & & 1 \end{pmatrix}.$$

Observe that the last column of T is just t_k of the above proof. A simple calculation shows that

$$T^{-1}A_nT = \begin{pmatrix} A_n^* & 0 \\ e_{k-1}^T & 1 \end{pmatrix}$$

where $e_{k-1}^T = (0, \dots, 0, 1)$ and

$$A_n^* = \begin{pmatrix} -\alpha_{k-2,n}^* & -\alpha_{k-3,n}^* & \dots & -\alpha_{1n}^* & -\alpha_{0n}^* \\ 1 & 0 & \dots & \cdot & 0 \\ & 1 & \dots & \cdot & 0 \\ & & \ddots & \vdots & \vdots \\ & & & 1 & 0 \end{pmatrix} \quad (5.21)$$

with

$$\begin{aligned} \alpha_{k-2,n}^* &= 1 + \alpha_{k-1,n}, & \alpha_{0n}^* &= -\alpha_{0n}, \\ \alpha_{k-j-1,n}^* - \alpha_{k-j,n}^* &= \alpha_{k-j,n} & \text{for } j &= 2, \dots, k-1. \end{aligned}$$

We remark that the coefficients $\alpha_{j,n}^*$ are just the coefficients of the polynomial defined by

$$\begin{aligned} &(\zeta^k + \alpha_{k-1,n}\zeta^{k-1} + \dots + \alpha_{1,n}\zeta + \alpha_{0,n}) \\ &= (\zeta - 1)(\zeta^{k-1} + \alpha_{k-2,n}\zeta^{k-2} + \dots + \alpha_{1,n}\zeta + \alpha_{0,n}^*). \end{aligned}$$

Theorem 5.6. *Let the method (5.15) be of order $p \geq 0$. Then the method is stable if and only if for all n and $l \geq 0$,*

$$\begin{aligned} a) \quad & \|A_{n+l}^* \cdots A_{n+1}^* A_n^*\| \leq M_1 \\ b) \quad & \|e_{k-1}^T \sum_{j=n}^{n+l} \prod_{i=n}^{j-1} A_i^*\| \leq M_2. \end{aligned}$$

Proof. A simple induction argument shows that

$$T^{-1} A_{n+l} \cdots A_n T = \begin{pmatrix} A_{n+l}^* \cdots A_n^* & 0 \\ b_{n,l}^T & 1 \end{pmatrix}$$

with

$$b_{n,l}^T = e_{k-1}^T \sum_{j=n}^{n+l} \prod_{i=n}^{j-1} A_i^*.$$

□

Since in this theorem the dimension of the matrices under consideration is reduced by one, it is especially useful for the stability investigation of two-step methods.

Example. Consider the two-step BDF-method (5.14). Here

$$\alpha_{0n} = \frac{\omega_{n+1}^2}{1 + 2\omega_{n+1}}, \quad \alpha_{1n} = -1 - \alpha_{0n}.$$

The matrix (5.21) becomes in this case

$$A_n^* = (-\alpha_{0n}^*), \quad -\alpha_{0n}^* = \frac{\omega_{n+1}^2}{1 + 2\omega_{n+1}}.$$

If $|\alpha_{0n}^*| \leq q < 1$ the conditions of Theorem 5.6 are satisfied and imply stability. This is the case, if

$$0 < h_{n+1}/h_n \leq \Omega < 1 + \sqrt{2}.$$

An interesting consequence of the theorem above is the *instability* of the two-step BDF-formula if the step sizes increase at least like $h_{n+1}/h_n \geq 1 + \sqrt{2}$.

The investigation of stability for k -step ($k \geq 3$) methods becomes much more difficult, because several step size ratios $\omega_{n+1}, \omega_{n+2}, \dots$ are involved. Grigorieff (1983) calculated the bounds (5.20) given in Table 5.1 for the higher order BDF-methods which *ensure* stability. These bounds are surely unrealistic, since all pathological step size variations are admitted.

A less pessimistic result is obtained if the step sizes are supposed to vary more smoothly (Gear & Tu 1974): the local error is known to be of the form $d(x_n)h_n^{p+1} + \mathcal{O}(h_n^{p+2})$, where $d(x)$ is the principal error function. This local error

Table 5.1. Bounds (5.20) for k -step BDF formulas

k	2	3	4	5
ω	0	0.836	0.979	0.997
Ω	2.414	1.127	1.019	1.003

is, by the step size control, kept equal to Tol . Hence, if $d(x)$ is bounded away from zero we have

$$h_n = |Tol/d(x_n)|^{1/(p+1)} + \mathcal{O}(h_n)$$

which implies (if $h_{n+1}/h_n \leq \Omega$) that

$$h_{n+1}/h_n = |d(x_n)/d(x_{n+1})|^{1/(p+1)} + \mathcal{O}(h_n).$$

If $d(x)$ is differentiable, we obtain

$$|h_{n+1}/h_n - 1| \leq Ch_n. \quad (5.22)$$

Several stability results of Gear & Tu are based on this hypothesis (“Consequently, we can expect either method to be stable if the fixed step method is stable. . .”). Adding up (5.22) we obtain

$$\sum_{j=n}^{n+l} |h_{j+1}/h_j - 1| \leq C(\hat{x} - x_0),$$

a condition which contains only step size ratios. This motivates the following theorem:

Theorem 5.7. *Let the coefficients α_{jn} of method (5.15) be continuously differentiable functions of $\omega_{n+1}, \dots, \omega_{n+k-1}$ in a neighbourhood of the set*

$$\{(\omega_{n+1}, \dots, \omega_{n+k-1}) ; \omega \leq \omega_j \leq \Omega\}$$

and assume that the method is stable for constant step sizes (i.e., for $\omega_j = 1$). Then the condition

$$\sum_{j=n}^{n+l} |h_{j+1}/h_j - 1| \leq C \quad \text{for all } n \text{ and } l \geq 0, \quad (5.23)$$

together with $\omega \leq h_{j+1}/h_j \leq \Omega$, imply the stability condition (5.19).

Proof. As in the proof of Theorem 5.5 we denote by A the companion matrix of the constant step size formula and by T a suitable transformation such that $\|T^{-1}AT\| = 1$. The mean value theorem, applied to $\alpha_j(\omega_{n+1}, \dots, \omega_{n+k-1}) - \alpha_j(1, \dots, 1)$, implies that

$$\|T^{-1}A_nT - T^{-1}AT\| \leq K \sum_{j=n+1}^{n+k-1} |\omega_j - 1|.$$

Hence

$$\|T^{-1}A_nT\| \leq 1 + K \sum_{j=n+1}^{n+k-1} |\omega_j - 1| \leq \exp\left(K \sum_{j=n+1}^{n+k-1} |\omega_j - 1|\right).$$

From this inequality we deduce that

$$\|A_{n+l} \dots A_{n+1} A_n\| \leq \|T\| \cdot \|T^{-1}\| \cdot \exp(K \cdot (k-1)C). \quad \square$$

Convergence

Convergence for variable step size Adams methods was first studied by Piotrowski (1969). In order to prove convergence for the general case we introduce the vector $Y_n = (y_{n+k-1}, \dots, y_{n+1}, y_n)^T$. In analogy to (4.8) the method (5.15) then becomes equivalent to

$$Y_{n+1} = (A_n \otimes I)Y_n + h_{n+k-1}\Phi_n(x_n, Y_n, h_n) \quad (5.24)$$

where A_n is given by (5.18) and

$$\Phi_n(x_n, Y_n, h_n) = (e_1 \otimes I)\Psi_n(x_n, Y_n, h_n).$$

The value $\Psi = \Psi_n(x_n, Y_n, h_n)$ is defined implicitly by

$$\Psi = \sum_{j=0}^{k-1} \beta_{jn} f(x_{n+j}, y_{n+j}) + \beta_{kn} f\left(x_{n+k}, h\Psi - \sum_{j=0}^{k-1} \alpha_{jn} y_{n+j}\right).$$

Let us further denote by

$$Y(x_n) = (y(x_{n+k-1}), \dots, y(x_{n+1}), y(x_n))^T$$

the exact values to be approximated by Y_n . The convergence theorem can now be formulated as follows:

Theorem 5.8. *Assume that*

- a) *the method (5.15) is stable, of order p , and has bounded coefficients α_{jn} and β_{jn} ;*
- b) *the starting values satisfy $\|Y(x_0) - Y_0\| = \mathcal{O}(h_0^p)$;*
- c) *the step size ratios are bounded ($h_n/h_{n-1} \leq \Omega$).*

Then the method is convergent of order p , i.e., for each differential equation $y' = f(x, y)$, $y(x_0) = y_0$ with f sufficiently differentiable the global error satisfies

$$\|y(x_n) - y_n\| \leq Ch^p \quad \text{for } x_n \leq \hat{x},$$

where $h = \max h_j$.

Proof. Since the method is of order p and the coefficients and step size ratios are bounded, formula (5.17) shows that the local error

$$\delta_{n+1} = Y(x_{n+1}) - (A_n \otimes I)Y(x_n) - h_{n+k-1}\Phi_n(x_n, Y(x_n), h_n) \quad (5.25)$$

satisfies

$$\delta_{n+1} = \mathcal{O}(h_n^{p+1}). \quad (5.26)$$

Subtracting (5.24) from (5.25) we obtain

$$\begin{aligned} Y(x_{n+1}) - Y_{n+1} &= (A_n \otimes I)(Y(x_n) - Y_n) \\ &\quad + h_{n+k-1}(\Phi_n(x_n, Y(x_n), h_n) - \Phi_n(x_n, Y_n, h_n)) + \delta_{n+1} \end{aligned}$$

and by induction it follows that

$$\begin{aligned} Y(x_{n+1}) - Y_{n+1} &= ((A_n \dots A_0) \otimes I)(Y(x_0) - Y_0) \\ &\quad + \sum_{j=0}^n h_{j+k-1}((A_n \dots A_{j+1}) \otimes I)(\Phi_j(x_j, Y(x_j), h_j) - \Phi_j(x_j, Y_j, h_j)) \\ &\quad + \sum_{j=0}^n ((A_n \dots A_{j+1}) \otimes I)\delta_{j+1}. \end{aligned}$$

As in the proof of Theorem 4.5 we deduce that the Φ_n satisfy a uniform Lipschitz condition with respect to Y_n . This, together with stability and (5.26), implies that

$$\|Y(x_{n+1}) - Y_{n+1}\| \leq \sum_{j=0}^n h_{j+k-1}L\|Y(x_j) - Y_j\| + C_1h^p.$$

In order to solve this inequality we introduce the sequence $\{\varepsilon_n\}$ defined by

$$\varepsilon_0 = \|Y(x_0) - Y_0\|, \quad \varepsilon_{n+1} = \sum_{j=0}^n h_{j+k-1}L\varepsilon_j + C_1h^p. \quad (5.27)$$

A simple induction argument shows that

$$\|Y(x_n) - Y_n\| \leq \varepsilon_n. \quad (5.28)$$

From (5.27) we obtain for $n \geq 1$

$$\varepsilon_{n+1} = \varepsilon_n + h_{n+k-1}L\varepsilon_n \leq \exp(h_{n+k-1}L)\varepsilon_n$$

so that also

$$\varepsilon_n \leq \exp((\hat{x} - x_0)L)\varepsilon_1 = \exp((\hat{x} - x_0)L) \cdot (h_{k-1}L\|Y(x_0) - Y_0\| + C_1h^p).$$

This inequality together with (5.28) completes the proof of Theorem 5.8. \square

Exercises

1. Prove that for constant step sizes the expressions $g_j(n)$ and $\Phi_j^*(n)$ (formulas (5.3) and (5.6)) reduce to

$$g_j(n) = \gamma_j, \quad \Phi_j^*(n) = \nabla^j f_n,$$

where γ_j is given by (1.6).

2. (Grigorieff 1983). For the k -step BDF-methods consider grids with constant mesh ratio ω , i.e., $h_n = \omega h_{n-1}$ for all n . In this case the elements of A_n^* (see (5.21)) are independent of n . Show numerically that all eigenvalues of A_n^* are of absolute value less than one for $0 < \omega < R_k$ where

k	2	3	4	5	6
R_k	2.414	1.618	1.280	1.127	1.044

III.6 Nordsieck Methods

While [the method] is primarily designed to optimize the efficiency of large-scale calculations on automatic computers, its essential procedures also lend themselves well to hand computation.
(A. Nordsieck 1962)

Two further problems must be dealt with in order to implement the automatic choice and revision of the elementary interval, namely, choosing which quantities to remember in such a way that the interval may be changed rapidly and conveniently . . .
(A. Nordsieck 1962)

In an important paper Nordsieck (1962) considered a class of methods for ordinary differential equations which allow a convenient way of changing the step size (see Section III.7). He already remarked that his methods are equivalent to the implicit Adams methods, in a certain sense. Let us begin with his derivation of these methods and then investigate their relation to linear multistep methods.

Nordsieck (1962) remarked “. . . that all methods of numerical integration are equivalent to finding an approximating polynomial for $y(x)$. . .”. His idea was to represent such a polynomial by the 0th to k th derivatives, i.e., by a vector (“the Nordsieck vector”)

$$z_n = \left(y_n, hy'_n, \frac{h^2}{2!}y''_n, \dots, \frac{h^k}{k!}y_n^{(k)} \right)^T. \quad (6.1)$$

The $y_n^{(j)}$ are meant to be approximations to $y^{(j)}(x_n)$, where $y(x)$ is the exact solution of the differential equation

$$y' = f(x, y). \quad (6.2)$$

In order to define the integration procedure we have to give a rule for determining z_{n+1} when z_n and the differential equation (6.2) are given. By Taylor's expansion, such a rule is (e.g., for $k = 3$)

$$\begin{aligned} y_{n+1} &= y_n + hy'_n + \frac{h^2}{2!}y''_n + \frac{h^3}{3!}y'''_n + \frac{h^4}{4!}e \\ hy'_{n+1} &= hy'_n + 2\frac{h^2}{2!}y''_n + 3\frac{h^3}{3!}y'''_n + 4\frac{h^4}{4!}e \\ \frac{h^2}{2!}y''_{n+1} &= \frac{h^2}{2!}y''_n + 3\frac{h^3}{3!}y'''_n + 6\frac{h^4}{4!}e \\ \frac{h^3}{3!}y'''_{n+1} &= \frac{h^3}{3!}y'''_n + 4\frac{h^4}{4!}e, \end{aligned} \quad (6.3)$$

where the value e is determined in such a way that

$$y'_{n+1} = f(x_{n+1}, y_{n+1}). \quad (6.4)$$

Inserting (6.4) into the second relation of (6.3) yields

$$4\frac{h^4}{4!}e = h \left(f(x_{n+1}, y_{n+1}) - f_n^p \right) \quad (6.5)$$

with

$$hf_n^p = hy'_n + 2\frac{h^2}{2!}y''_n + 3\frac{h^3}{3!}y'''_n.$$

With this relation for e the above method becomes

$$\begin{aligned} y_{n+1} &= y_n + hy'_n + \frac{h^2}{2!}y''_n + \frac{h^3}{3!}y'''_n + \frac{1}{4}h\left(f(x_{n+1}, y_{n+1}) - f_n^p\right) \\ hy'_{n+1} &= hy'_n + 2\frac{h^2}{2!}y''_n + 3\frac{h^3}{3!}y'''_n + h\left(f(x_{n+1}, y_{n+1}) - f_n^p\right) \\ \frac{h^2}{2!}y''_{n+1} &= \frac{h^2}{2!}y''_n + 3\frac{h^3}{3!}y'''_n + \frac{3}{2}h\left(f(x_{n+1}, y_{n+1}) - f_n^p\right) \\ \frac{h^3}{3!}y'''_{n+1} &= \frac{h^3}{3!}y'''_n + h\left(f(x_{n+1}, y_{n+1}) - f_n^p\right) \end{aligned} \quad (6.6)$$

The first equation constitutes an implicit formula for y_{n+1} , the others are explicit. Observe that for sufficiently accurate approximations $y_n^{(j)}$ to $y^{(j)}(x_n)$ the value e (formula (6.5)) is an approximation to $y^{(4)}(x_n)$. This seems to be a desirable property from the point of view of accuracy. Unfortunately, method (6.6) is unstable. To see this, we put $f(x, y) = 0$ in (6.6). In this case the method becomes the linear transformation

$$z_{n+1} = Mz_n \quad (6.7)$$

where

$$M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1/4 \\ 1 \\ 3/2 \\ 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 2 & 3 \end{pmatrix}.$$

The eigenvalues of M are seen to be $1, 0, -(2 + \sqrt{3})$ and $-1/(2 + \sqrt{3})$, implying that (6.6) is unstable and therefore of no use. The phenomenon that highly accurate methods are often unstable is, after our experiences in Section III.3, no longer astonishing.

To overcome this difficulty Nordsieck proposed to replace the constants $1/4, 1, 3/2, 1$ which appear in front of the brackets in (6.6) by arbitrary values (l_0, l_1, l_2, l_3) , and to use this extra freedom to achieve stability. In compact form this modification can be written as

$$z_{n+1} = (P \otimes I)z_n + (l \otimes I)\left(hf(x_{n+1}, y_{n+1}) - (e_1^T P \otimes I)z_n\right). \quad (6.8)$$

Here z_n is given by (6.1), P is the Pascal triangle matrix defined by

$$p_{ij} = \begin{cases} \binom{j}{i} & \text{for } 0 \leq i \leq j \leq k, \\ 0 & \text{else,} \end{cases}$$

$l = (l_0, l_1, \dots, l_k)^T$ and $e_1 = (0, 1, 0, \dots, 0)^T$. Observe that the indices of vectors and matrices start from zero.

For notational simplicity in the following theorems, we consider from now on scalar differential equations only, so that method (6.8) becomes

$$z_{n+1} = Pz_n + l(hf_{n+1} - e_1^T Pz_n). \quad (6.8')$$

All results, of course, remain valid for systems of equations. Condition (6.4), which relates the method to the differential equation, fixes the value of l_1 as

$$l_1 = 1. \quad (6.9)$$

The above stability analysis applied to the general method (6.8) leads to the difference equation (6.7) with

$$M = P - le_1^T P. \quad (6.10)$$

For instance, for $k = 3$ this matrix is given by

$$M = \begin{pmatrix} 1 & 1 - l_0 & 1 - 2l_0 & 1 - 3l_0 \\ 0 & 0 & 0 & 0 \\ 0 & -l_2 & 1 - 2l_2 & 3 - 3l_2 \\ 0 & -l_3 & -2l_3 & 1 - 3l_3 \end{pmatrix}.$$

One observes that 1 and 0 are two eigenvalues of M and that its characteristic polynomial is independent of l_0 . Nordsieck determined l_2, \dots, l_k in such a way that the remaining eigenvalues of M are zero. For $k = 3$ this yields $l_2 = 3/4$ and $l_3 = 1/6$. The coefficient l_0 can be chosen such that the error constant of the method (see Theorem 6.2 below) vanishes. In our situation one gets $l_0 = 3/8$, so that the resulting method is given by

$$l = (3/8, 1, 3/4, 1/6)^T.$$

It is interesting to note that this method is equivalent to the implicit 3-step Adams method. Indeed, an elimination of the terms $(h^3/3!)y_n'''$ and $(h^2/2!)y_n''$ by using formula (6.8) with reduced indices leads to (cf. formula (1.9'))

$$y_{n+1} = y_n + \frac{h}{24} (9y'_{n+1} + 19y'_n - 5y'_{n-1} + y'_{n-2}). \quad (6.11)$$

Equivalence with Multistep Methods

More insight into the connection between Nordsieck methods and multistep methods is due to Descloux (1963), Osborne (1966), and Skeel (1979). The following two theorems show that every Nordsieck method is equivalent to a multistep formula and that the order of this method is at least k .

Theorem 6.1. Consider the Nordsieck method (6.8) where $l_1 = 1$. The first two components of z_n then satisfy the linear multistep formula (for $n \geq 0$)

$$\sum_{i=0}^k \alpha_i y_{n+i} = h \sum_{i=0}^k \beta_i f_{n+i} \quad (6.12)$$

where the generating polynomials are given by

$$\begin{aligned} \varrho(\zeta) &= \det(\zeta I - P) \cdot e_1^T (\zeta I - P)^{-1} l \\ \sigma(\zeta) &= \det(\zeta I - P) \cdot e_0^T (\zeta I - P)^{-1} l. \end{aligned} \quad (6.13)$$

Proof. The proof of the original papers simplifies considerably, if we work with the generating functions (discrete Laplace transformation)

$$Z(\zeta) = \sum_{n \geq 0} z_n \zeta^n, \quad Y(\zeta) = \sum_{n \geq 0} y_n \zeta^n, \quad F(\zeta) = \sum_{n \geq 0} f_n \zeta^n, \quad \dots$$

Multiplying formula (6.8') by ζ^{n+1} and adding up we obtain

$$Z(\zeta) = \zeta P Z(\zeta) + l \left(h F(\zeta) - e_1^T P \zeta Z(\zeta) \right) + (z_0 - l h f_0). \quad (6.14)$$

Similarly, the linear multistep method (6.12) can be written as

$$\widehat{\varrho}(\zeta) Y(\zeta) = h \widehat{\sigma}(\zeta) F(\zeta) + p_{k-1}(\zeta), \quad (6.15)$$

where

$$\widehat{\varrho}(\zeta) = \zeta^k \varrho(1/\zeta), \quad \widehat{\sigma}(\zeta) = \zeta^k \sigma(1/\zeta) \quad (6.16)$$

and p_{k-1} is a polynomial of degree $k-1$ depending on the starting values. In order to prove the theorem we have to show that the first two components of $Z(\zeta)$ satisfy a relation of the form (6.15). We first rewrite equation (6.14) in the form

$$Z(\zeta) = (I - \zeta P)^{-1} l \left(h F(\zeta) - e_1^T P \zeta Z(\zeta) \right) + (I - \zeta P)^{-1} (z_0 - l h f_0)$$

so that its first two components become

$$\begin{aligned} Y(\zeta) &= e_0^T (I - \zeta P)^{-1} l \left(h F(\zeta) - e_1^T P \zeta Z(\zeta) \right) + e_0^T (I - \zeta P)^{-1} (z_0 - l h f_0) \\ h F(\zeta) &= e_1^T (I - \zeta P)^{-1} l \left(h F(\zeta) - e_1^T P \zeta Z(\zeta) \right) + e_1^T (I - \zeta P)^{-1} (z_0 - l h f_0). \end{aligned}$$

Eliminating the term in brackets and multiplying by $\det(I - \zeta P)$ we arrive at formula (6.15) with

$$\begin{aligned} \widehat{\varrho}(\zeta) &= \det(I - \zeta P) \cdot e_1^T (I - \zeta P)^{-1} l \\ \widehat{\sigma}(\zeta) &= \det(I - \zeta P) \cdot e_0^T (I - \zeta P)^{-1} l \\ p_{k-1}(\zeta) &= \det(I - \zeta P) \left(e_1^T (I - \zeta P)^{-1} l e_0^T (I - \zeta P)^{-1} \right. \\ &\quad \left. - e_0^T (I - \zeta P)^{-1} l e_1^T (I - \zeta P)^{-1} \right) z_0. \end{aligned} \quad (6.17)$$

With the help of (6.16) we immediately get formulas (6.13). Therefore, it remains to show that p_{k-1} , given by (6.17), is a polynomial of degree $k-1$. Since the dimension of P is $(k+1)$, p_{k-1} behaves like ζ^{k-1} for $|\zeta| \rightarrow \infty$. Finally, the relation (6.15) implies that the Laurent series of p_{k-1} cannot contain negative powers. \square

Putting $(\zeta I - P)^{-1}l = u$ in (6.13) and applying Cramer's rule to the linear system $(\zeta I - P)u = l$ we obtain from (6.13) the elegant expressions

$$\varrho(\zeta) = \det \begin{pmatrix} \zeta - 1 & l_0 & -1 & \dots & -1 \\ 0 & l_1 & -2 & \dots & -k \\ 0 & l_2 & \zeta - 1 & \dots & \cdot \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & l_k & 0 & \dots & \zeta - 1 \end{pmatrix} \quad (6.13a)$$

$$\sigma(\zeta) = \det \begin{pmatrix} l_0 & -1 & -1 & \dots & -1 \\ l_1 & \zeta - 1 & -2 & \dots & -k \\ l_2 & 0 & \zeta - 1 & \dots & \cdot \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_k & 0 & 0 & \dots & \zeta - 1 \end{pmatrix}. \quad (6.13b)$$

We observe that $\varrho(\zeta)$ does not depend on l_0 . Further, $\zeta_0 = 1$ is a simple root of $\varrho(\zeta)$ if and only if $l_k \neq 0$. We have

$$\varrho'(1) = \sigma(1) = k! l_k. \quad (6.18)$$

Condition (6.9) is equivalent to $\alpha_k = 1$.

Theorem 6.2. Assume that $l_k \neq 0$. The multistep method defined by (6.13) is of order at least k and its error constant (see (2.13)) is given by

$$C = -\frac{b^T l}{k! l_k}.$$

Here the components of

$$b^T = (B_0, B_1, \dots, B_k) = \left(1, -\frac{1}{2}, \frac{1}{6}, 0, -\frac{1}{30}, 0, \frac{1}{42}, \dots\right)$$

are the Bernoulli numbers.

Proof. By Theorem 2.4 we have order k iff

$$\varrho(\zeta) - \log \zeta \cdot \sigma(\zeta) = C_{k+1}(\zeta - 1)^{k+1} + \mathcal{O}((\zeta - 1)^{k+2}).$$

Since $\det(\zeta I - P) = (\zeta - 1)^{k+1}$ this is equivalent to

$$e_1^T (\zeta I - P)^{-1} l - \log \zeta \cdot e_0^T (\zeta I - P)^{-1} l = C_{k+1} + \mathcal{O}((\zeta - 1))$$

and, by (6.18), it suffices to show that

$$(\log \zeta \cdot e_0^T - e_1^T)(\zeta I - P)^{-1} = b^T + \mathcal{O}((\zeta - 1)). \quad (6.19)$$

Denoting the left-hand side of (6.19) by $b^T(\zeta)$ we obtain

$$(\zeta I - P)^T b(\zeta) = (\log \zeta \cdot e_0 - e_1). \quad (6.20)$$

The q th component ($q \geq 2$) of this equation

$$\zeta b_q(\zeta) - \sum_{j=0}^q \binom{q}{j} b_j(\zeta) = 0$$

is equivalent to

$$\frac{\zeta b_q(\zeta)}{q!} - \sum_{j=0}^q \frac{b_j(\zeta)}{j!} \frac{1}{(q-j)!} = 0,$$

which is seen to be a Cauchy product. Hence, formula (6.20) becomes

$$\zeta \sum_{q \geq 0} \frac{t^q}{q!} b_q(\zeta) - e^t \sum_{q \geq 0} \frac{t^q}{q!} b_q(\zeta) = \log \zeta - t$$

which yields

$$\sum_{q \geq 0} \frac{t^q}{q!} b_q(\zeta) = \frac{t - \log \zeta}{e^t - \zeta}.$$

If we set $\zeta = 1$ in this formula we obtain

$$\sum_{q \geq 0} \frac{t^q}{q!} b_q(1) = \frac{t}{e^t - 1},$$

therefore $b_q(1) = B_q$, the q th Bernoulli number (see Abramowitz & Stegun, Chapter 23). \square

We have thus shown that to each Nordsieck method (6.8) there corresponds a linear multistep method of order at least k . Our next aim is to establish a correspondence in the opposite direction.

Theorem 6.3. *Let (ϱ, σ) be the generating polynomials of a k -step method (6.12) of order at least k and assume $\alpha_k = 1$. Then we have:*

- a) *There exists a unique vector l such that ϱ and σ are given by (6.13).*
- b) *If, in addition, the multistep method is irreducible, then there exists a non-singular transformation T such that the solution of (6.8') is related to that of (6.12) by*

$$z_n = T^{-1} u_n \quad (6.21)$$

where the j th component of u_n is given by

$$u_j^{(n)} = \begin{cases} \sum_{i=0}^j (\alpha_{k-j+i} y_{n+i} - h\beta_{k-j+i} f_{n+i}) & \text{for } 0 \leq j \leq k-1, \\ hf_n & \text{for } j = k. \end{cases} \quad (6.22)$$

Proof. a) For every k th order multistep method the polynomial $\varrho(\zeta)$ is uniquely determined by $\sigma(\zeta)$ (see Theorem 2.4). Expanding the determinant in (6.13b) with respect to the first column we see that

$$\sigma(\zeta) = l_0(\zeta - 1)^k + l_1(\zeta - 1)^{k-1}r_1(\zeta) + \dots + l_k r_k(\zeta),$$

where $r_j(\zeta)$ is a polynomial of degree j satisfying $r_j(1) \neq 0$. Hence, l can be computed from $\sigma(\zeta)$.

b) Let y_0, \dots, y_{k-1} and f_0, \dots, f_{k-1} be given. Then the polynomial $p_{k-1}(\zeta)$ in (6.15) satisfies

$$p_{k-1}(\zeta) = u_0^{(0)} + u_1^{(0)}\zeta + \dots + u_{k-1}^{(0)}\zeta^{k-1}.$$

On the other hand, if the starting vector z_0 for the Nordsieck method defined by l of (a) is known, then $p_{k-1}(\zeta)$ is given by (6.17). Equating both expressions we obtain

$$\sum_{j=0}^{k-1} u_j^{(0)} \zeta^j = (\widehat{\varrho}(\zeta)e_0^T - \widehat{\sigma}(\zeta)e_1^T)(I - \zeta P)^{-1}z_0. \quad (6.23)$$

We now denote by t_j^T ($j = 0, \dots, k-1$) the coefficients of the vector polynomial

$$(\widehat{\varrho}(\zeta)e_0^T - \widehat{\sigma}(\zeta)e_1^T)(I - \zeta P)^{-1} = \sum_{j=0}^{k-1} t_j^T \zeta^j \quad (6.24)$$

and set $t_k^T = e_1^T$. Then let T be the square matrix whose j th row is t_j^T so that $u_0 = Tz_0$ is a consequence of (6.23) and $hf_n = hy'_n$. The same argument applied to y_n, \dots, y_{n+k-1} and f_n, \dots, f_{n+k-1} instead of y_0, \dots, y_{k-1} and f_0, \dots, f_{k-1} yields $u_n = Tz_n$ for all n .

To complete the proof it remains to verify the non-singularity of T . Let $v = (v_0, v_1, \dots, v_k)^T$ be a non-zero vector satisfying $Tv = 0$. By definition of t_k^T we have $v_1 = 0$ and from (6.24) it follows (using the transformation (6.16)) that

$$\varrho(\zeta)\tau_0(\zeta) = \sigma(\zeta)\tau_1(\zeta), \quad (6.25)$$

where $\tau_i(\zeta) = \det(\zeta I - P)e_i^T(\zeta I - P)^{-1}v$ are polynomials of degree at most k . Moreover, Cramer's rule shows that the degree of $\tau_1(\zeta)$ is at most $k-1$, since $v_1 = 0$. Hence from (6.25) at least one of the roots of $\varrho(\zeta)$ must be a root of $\sigma(\zeta)$. This is in contradiction with the assumption that the method is irreducible. \square

Table 6.1. Coefficients l_j of the k -step implicit Adams methods

	l_0	l_1	l_2	l_3	l_4	l_5	l_6
$k = 1$	1/2	1					
$k = 2$	5/12	1	1/2				
$k = 3$	3/8	1	3/4	1/6			
$k = 4$	251/720	1	11/12	1/3	1/24		
$k = 5$	95/288	1	25/24	35/72	5/48	1/120	
$k = 6$	19087/60480	1	137/120	5/8	17/96	1/40	1/720

Table 6.2. Coefficients l_j of the k -step BDF-methods

	l_0	l_1	l_2	l_3	l_4	l_5	l_6
$k = 1$	1	1					
$k = 2$	2/3	1	1/3				
$k = 3$	6/11	1	6/11	1/11			
$k = 4$	12/25	1	7/10	1/5	1/50		
$k = 5$	60/137	1	225/274	85/274	15/274	1/274	
$k = 6$	20/49	1	58/63	5/12	25/252	1/84	1/1764

The vectors l which correspond to the implicit Adams methods and to the BDF-methods are given in Tables 6.1 and 6.2. For these two classes of methods we shall investigate the equivalence in some more detail.

Implicit Adams Methods

The following results are due to Byrne & Hindmarsh (1975). Since their “efficient package” EPISODE and the successor VODE are based on the Nordsieck representation of variable step size methods, we extend our considerations to this case. The Adams methods define in a natural way a polynomial which approximates the unknown solution of (6.2). Namely, if y_n and f_n, \dots, f_{n-k+1} are given, then the k -step Adams method is equivalent to the construction of a polynomial $p_{n+1}(x)$ of degree $k+1$ which satisfies

$$\begin{aligned} p_{n+1}(x_n) &= y_n, & p_{n+1}(x_{n+1}) &= y_{n+1}, \\ p'_{n+1}(x_j) &= f_j & \text{for } j &= n-k+1, \dots, n+1. \end{aligned} \quad (6.26)$$

Condition (6.26) defines y_{n+1} implicitly. We observe that the difference of two consecutive polynomials, $p_{n+1}(x) - p_n(x)$, vanishes at x_n and that its derivative

is zero at x_{n-k+1}, \dots, x_n . Therefore, if we let $e_{n+1} = y_{n+1} - p_n(x_{n+1})$, this difference can be written as

$$p_{n+1}(x) - p_n(x) = \Lambda \left(\frac{x - x_{n+1}}{x_{n+1} - x_n} \right) e_{n+1} \quad (6.27)$$

where Λ is the unique polynomial of degree $(k+1)$ defined by

$$\begin{aligned} \Lambda(0) &= 1, & \Lambda(-1) &= 0 \\ \Lambda' \left(\frac{x_j - x_{n+1}}{x_{n+1} - x_n} \right) &= 0 & \text{for } j &= n-k+1, \dots, n. \end{aligned} \quad (6.28)$$

The derivative of (6.27) taken at $x = x_{n+1}$ shows that with $h_n = x_{n+1} - x_n$,

$$h_n f_{n+1} - h_n p'_n(x_{n+1}) = \Lambda'(0) e_{n+1}.$$

If we introduce the Nordsieck vector

$$\tilde{z}_n = \left(p_n(x_n), h_n p'_n(x_n), \dots, \frac{h_n^{k+1}}{(k+1)!} p_n^{(k+1)}(x_n) \right)^T$$

and the coefficients \tilde{l}_j by

$$\Lambda(t) = \sum_{j=0}^{k+1} \tilde{l}_j t^j, \quad (6.29)$$

then (6.27) becomes equivalent to

$$\tilde{z}_{n+1} = P \tilde{z}_n + \tilde{l} \tilde{l}_1^{-1} (h f_{n+1} - e_1^T P \tilde{z}_n) \quad (6.30)$$

with $\tilde{l} = (\tilde{l}_0, \tilde{l}_1, \dots, \tilde{l}_{k+1})^T$. This method is of the form (6.8'). However, it is of dimension $k+2$ and not, as expected by Theorem 6.3, of dimension $k+1$. The reason is the following: let $\tilde{\varrho}(\zeta)$ and $\tilde{\sigma}(\zeta)$ be the generating polynomials of the multistep method which corresponds to (6.30). Then the conditions $\Lambda(-1) = 0$ and $\Lambda'(-1) = 0$ imply that $\tilde{\sigma}(0) = \tilde{\varrho}(0) = 0$, so that this method is reducible. Nevertheless, method (6.30) is useful, since the last component of \tilde{z}_n can be used for step size control.

Remark. For $k \geq 2$ the coefficients \tilde{l}_j , defined by (6.29), depend on the step size ratios h_j/h_{j-1} for $j = n-k+2, \dots, n$. They can be computed from the formula

$$\Lambda(t) = \frac{\int_{-1}^t \prod_{j=1}^k (s - t_j) ds}{\int_{-1}^0 \prod_{j=1}^k (s - t_j) ds} \quad (6.31)$$

where $t_j = (x_{n-j+1} - x_{n+1}) / (x_{n+1} - x_n)$ (see also Exercise 1).

BDF-Methods

One step of the k -step BDF method consists in constructing a polynomial $q_{n+1}(x)$ of degree k which satisfies

$$\begin{aligned} q_{n+1}(x_j) &= y_j & \text{for } j = n - k + 1, \dots, n + 1 \\ q'_{n+1}(x_{n+1}) &= f_{n+1} \end{aligned} \quad (6.32)$$

and in computing a value y_{n+1} which makes this possible. As for the Adams methods we have

$$q_{n+1}(x) - q_n(x) = \Lambda\left(\frac{x - x_{n+1}}{x_{n+1} - x_n}\right) \cdot (y_{n+1} - q_n(x_{n+1})), \quad (6.33)$$

where $\Lambda(t)$ is the polynomial of degree k defined by

$$\begin{aligned} \Lambda\left(\frac{x_j - x_{n+1}}{x_{n+1} - x_n}\right) &= 0 & \text{for } j = n - k + 1, \dots, n, \\ \Lambda(0) &= 1. \end{aligned}$$

With the vector

$$\tilde{z}_n = \left(q_n(x_n), h_n q'_n(x_n), \dots, \frac{h_n^k}{k!} q_n^{(k)}(x_n) \right)^T$$

and the coefficients \tilde{l}_j given by

$$\Lambda(t) = \sum_{j=0}^k \tilde{l}_j t^j,$$

equation (6.33) becomes

$$\tilde{z}_{n+1} = P \tilde{z}_n + \tilde{l} \tilde{l}_1^{-1} (h f_{n+1} - e_1^T P \tilde{z}_n). \quad (6.34)$$

The vector $\tilde{l} = (\tilde{l}_0, \tilde{l}_1, \dots, \tilde{l}_k)^T$ can be computed from the formula

$$\Lambda(t) = \prod_{j=1}^k \left(1 + \frac{t}{t_j} \right)$$

where $t_j = (x_{n-j+1} - x_{n+1}) / (x_{n+1} - x_n)$. For constant step sizes formula (6.34) corresponds to that of Theorem 6.3 and the coefficients $l_j = \tilde{l}_j / \tilde{l}_1$ coincide with those of Table 6.2.

Exercises

1. Let $l_j^{(k)} (j = 0, \dots, k)$ be the Nordsieck coefficients of the k -step implicit Adams methods (defined by Theorem 6.3 and given in Table 6.1). Further, denote by $\tilde{l}_j^{(k)} (j = 0, \dots, k+1)$ the coefficients given by (6.29) and (6.31) for the case of constant step sizes. Show that

$$\frac{\tilde{l}_j^{(k)}}{\tilde{l}_1^{(k)}} = \begin{cases} l_j^{(k)} & \text{for } j = 0 \\ l_j^{(k+1)} & \text{for } j = 1, \dots, k+1. \end{cases}$$

Use these relations to verify Table 6.1.

2. a) Calculate the matrix T of Theorem 6.3 for the 3-step implicit Adams method.

Result.

$$T^{-1} = \begin{pmatrix} 1 & 0 & 0 & 3/8 \\ 0 & 0 & 0 & 1 \\ 0 & 6 & 6 & 3/4 \\ 0 & 4 & 12 & 1/6 \end{pmatrix}.$$

Show that the Nordsieck vector z_n is given by

$$z_n = \left(y_n, hf_n, (3hf_n - 4hf_{n-1} + hf_{n-2})/4, (hf_n - 2hf_{n-1} + hf_{n-2})/6 \right)^T.$$

- b) The vector \tilde{z}_n for the 2-step implicit Adams method (6.30) (constant step sizes) also satisfies

$$\tilde{z}_n = \left(y_n, hf_n, (3hf_n - 4hf_{n-1} + hf_{n-2})/4, (hf_n - 2hf_{n-1} + hf_{n-2})/6 \right)^T,$$

but this time y_n is a less accurate approximation to $y(x_n)$.

III.7 Implementation and Numerical Comparisons

There is a great deal of freedom in the implementation of multistep methods (even if we restrict our considerations to the Adams methods). One can either directly use the *variable step size methods* of Section III.5 or one can take a fixed step size method and determine the necessary offgrid values, which are needed for a change of step size, by *interpolation*. Further, it is possible to choose between the *divided difference* formulation (5.7) and the *Nordsieck* representation (6.30).

The historical approach was the use of formula (1.9) together with interpolation (J.C. Adams (1883): “We may, of course, change the value of ω (the step size) whenever the more or less rapid rate of diminution of the successive differences shews that it is expedient to increase or diminish the interval. It is only necessary, by selection from or interpolation between the values already calculated, to find the coordinates for a few values of φ separated from each other by the newly chosen interval.”). It is theoretically more satisfactory and more elegant to work with the variable step size method (5.7). For both of these approaches the change of step size is rather expensive whereas the change of order is very simple — one just has to add a further term to the expansion (1.9). If the Nordsieck representation (6.30) is implemented, the situation is the opposite. There, the change of order is not as direct as above, but the step size can be changed simply by multiplying the Nordsieck-vector (6.1) by the diagonal matrix with entries $(1, \omega, \omega^2, \dots)$ where $\omega = h_{\text{new}}/h_{\text{old}}$ is the step size ratio. Indeed, this was the main reason for introducing this representation.

Step Size and Order Selection

Much was made of the starting of multistep computations and the need for Runge-Kutta methods in the literature of the 60ies (see e.g., Ralston 1962). Nowadays, codes for multistep methods simply start with order one and very small step sizes and are therefore self-starting. The following step size and order selection is closely related to the description of Shampine & Gordon (1975).

Suppose that the numerical integration has proceeded successfully until x_n and that a further step with step size h_n and order $k+1$ is taken, which yields the

approximation y_{n+1} to $y(x_{n+1})$. To decide whether y_{n+1} will be accepted or not, we need an estimate of the local truncation error. Such an estimate is e.g. given by

$$le_{k+1}(n+1) = y_{n+1}^* - y_{n+1}$$

where y_{n+1}^* is the result of the $(k+2)$ nd order implicit Adams formula. Subtracting formula (5.7) from the same formula with k replaced by $k+1$, we obtain

$$le_{k+1}(n+1) = h_n(g_{k+1}(n) - g_k(n))\Phi_{k+1}(n+1). \quad (7.1)$$

Without changing the leading term in this expression we can replace the expression $\Phi_{k+1}(n+1)$ by

$$\Phi_{k+1}^p(n+1) = \prod_{i=0}^k (x_{n+1} - x_{n-i}) \delta^{k+1} f^p[x_{n+1}, x_n, \dots, x_{n-k}]. \quad (7.2)$$

The superscript p of f indicates that $f_{n+1} = f(x_{n+1}, y_{n+1})$ is replaced by $f(x_{n+1}, p_{n+1})$ when forming the divided differences. If the implicit equation (5.7) is solved iteratively with p_{n+1} as predictor, then $\Phi_{k+1}^p(n+1)$ has to be calculated anyway. Therefore, the only cost for computing the estimate

$$LE_{k+1}(n+1) = h_n(g_{k+1}(n) - g_k(n))\Phi_{k+1}^p(n+1) \quad (7.3)$$

is the computation of $g_{k+1}(n)$. After the expression (7.3) has been calculated, we require (in the norm (4.11) of Section II.4)

$$\|LE_{k+1}(n+1)\| \leq 1 \quad (7.4)$$

for the step to be successful.

If the Nordsieck representation (6.30) is considered instead of (5.7), then the estimate of the local error is not as simple, since the \tilde{l} -vectors in (6.30) are totally different for different orders. For a possible error-estimate we refer to the article of Byrne & Hindmarsh (1975).

Suppose now that y_{n+1} is accepted. We next have to choose a new step size and a new order. The idea of the *step size selection* is to find the largest h_{n+1} for which the predicted local error is acceptable, i.e., for which

$$h_{n+1} \cdot |g_{k+1}(n+1) - g_k(n+1)| \cdot \|\Phi_{k+1}^p(n+2)\| \leq 1.$$

However, this procedure is of no practical use, since the expressions $g_j(n+1)$ and $\Phi_{k+1}^p(n+2)$ depend in a complicated manner on the unknown step size h_{n+1} . Also, the coefficients $g_{k+1}(n+1)$ and $g_k(n+1)$ are too expensive to calculate. To overcome this difficulty we assume the grid to be equidistant (this is a doubtful assumption, but leads to a simple formula for the new step size). In this case the local error (for the method of order $k+1$) is of the form $C(x_{n+2})h^{k+2} + \mathcal{O}(h^{k+3})$ with C depending smoothly on x . The local error at x_{n+2} can thus be approximated by that at x_{n+1} and in the same way as for one-step methods (cf. Section II.4

formula (4.12)) we obtain

$$h_{\text{opt}}^{(k+1)} = h_n \cdot \left(\frac{1}{\|LE_{k+1}(n+1)\|} \right)^{1/(k+2)} \quad (7.5)$$

as optimal step size. The local error $LE_{k+1}(n+1)$ is given by (7.3) or, again under the assumption of an equidistant grid, by

$$LE_{k+1}(n+1) = h_n \gamma_{k+1}^* \Phi_{k+1}^p(n+1) \quad (7.6)$$

with γ_{k+1}^* from Table 1.2 (see Exercise 1 of Section III.5 and Exercise 4 of Section III.1).

We next describe how an *optimal order* can be determined. Since the number of necessary function evaluations is the same for all orders, there are essentially two strategies for selecting the new order. One can choose the order $k+1$ either such that the local error estimate is minimal, or such that the new optimal step size is maximal. Because of the exponent $1/(k+2)$ in formula (7.5), the two strategies are not always equivalent. For more details see the description of the code DEABM below. It should be mentioned that each implementation of the Adams methods — and there are many — contains refinements of the above description and has in addition several ad-hoc devices. One of them is to keep the step size constant if $h_{\text{new}}/h_{\text{old}}$ is near to 1. In this way the computation of the coefficients $g_j(n)$ is simplified.

Some Available Codes

We have chosen the three codes DEABM, VODE and LSODE to illustrate the order- and step size strategies for multistep methods.

DEABM is a modification of the code DE/STEP/INTRP described in the book of Shampine & Gordon (1975). It belongs to the package DEPAC, designed by Shampine & Watts (1979). Our numerical tests use the revised version from February 1984. For European users it is available from the “Rechenzentrum der RWTH Aachen, Seffenter Weg 23, D-5100 Aachen, Germany”.

This code implements the variable step size, divided difference representation (5.7) of the Adams formulas. In order to solve the nonlinear equation (5.7) for y_{n+1} the value p_{n+1} is taken as predictor (P), then $f_{n+1}^p = f(x_{n+1}, p_{n+1})$ is calculated (E) and *one* corrector iteration (C) is performed, to obtain y_{n+1} . Finally, in the case of a successful step, $f_{n+1} = f(x_{n+1}, y_{n+1})$ is evaluated (E) for the next step. This PECE implementation needs two function evaluations for each successful step. Let us also outline the order strategy of this code: after performing a step with order $k+1$, one computes $LE_{k-1}(n+1)$, $LE_k(n+1)$ and $LE_{k+1}(n+1)$ using a slight modification of (7.6). Then the order is reduced by one, if

$$\max\left(\|LE_{k-1}(n+1)\|, \|LE_k(n+1)\|\right) \leq \|LE_{k+1}(n+1)\|. \quad (7.7)$$

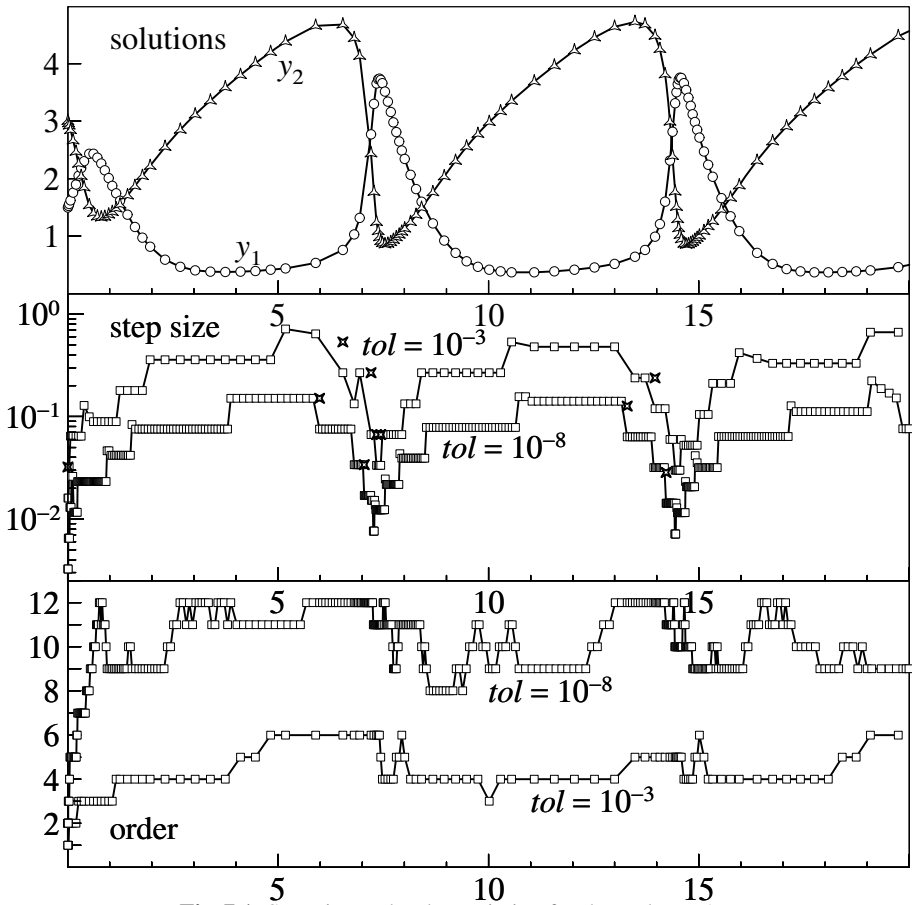


Fig. 7.1. Step size and order variation for the code DEABM

An increase in the order is considered only if the step is successful, (7.7) is violated and a constant step size is used. In this case one computes the estimate

$$LE_{k+2}(n+1) = h_n \gamma_{k+2}^* \Phi_{k+2}(n+1)$$

using the new value $f_{n+1} = f(x_{n+1}, y_{n+1})$ and increases the order by one if

$$\|LE_{k+2}(n+1)\| < \|LE_{k+1}(n+1)\|.$$

In Fig. 7.1 we demonstrate the variation of the step size and order on the example of Section II.4 (see Fig. 4.1 and also Fig. 9.5 of Section II.9). We plot the solution obtained with $Rtol = Atol = 10^{-3}$, the step size and order for the tolerances 10^{-3} and 10^{-8} . We observe that the step size — and not the order — drops significantly at passages where the solution varies more rapidly. Furthermore, constant step sizes are taken over long intervals, and the order is changed rather often (especially for $Tol = 10^{-8}$). This is in agreement with the observation of Shampine &

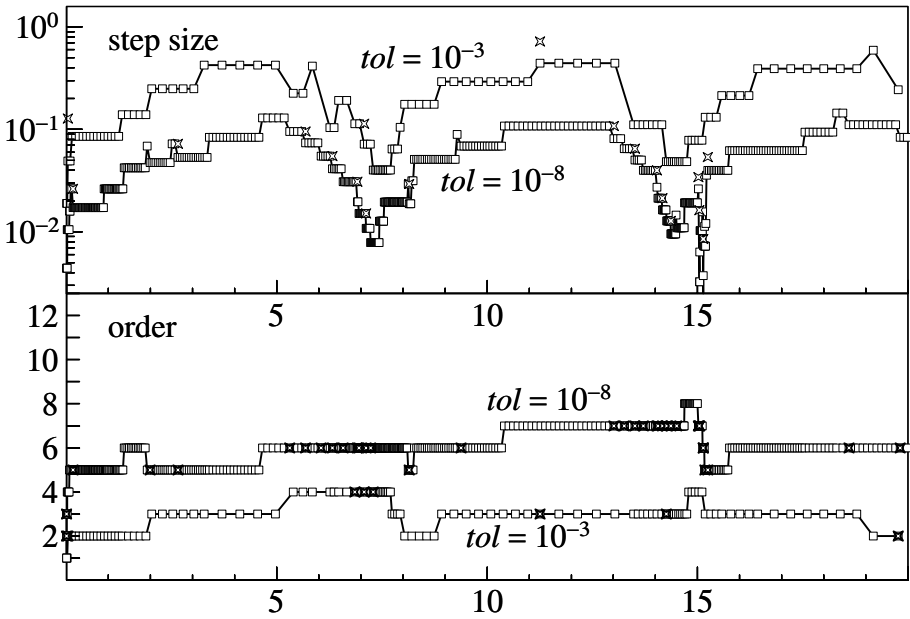


Fig. 7.2. Step size and order variation for the code VODE

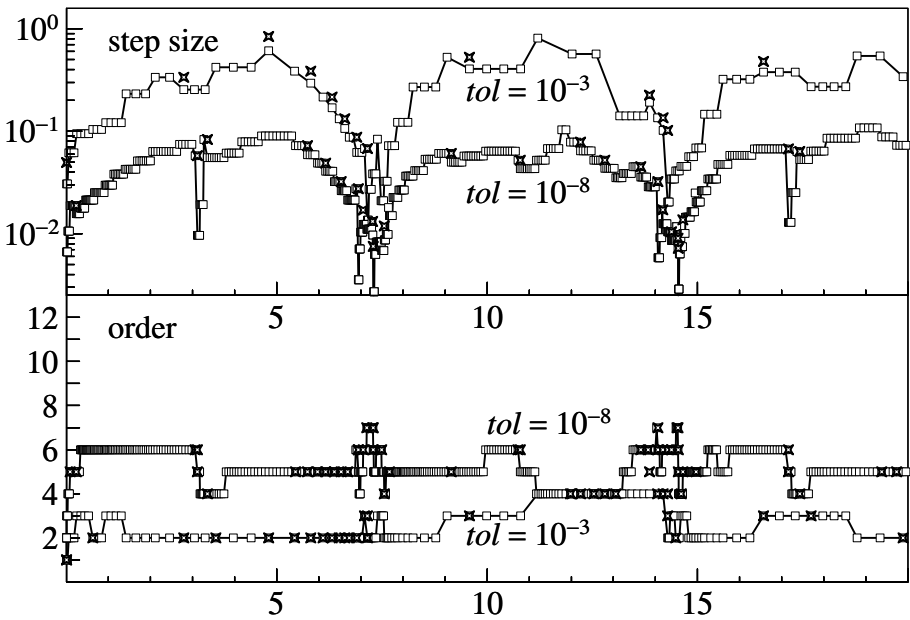


Fig. 7.3. Step size and order variation for the code LSODE

Gordon (1975): "... small reductions in the estimated error may cause the order to fluctuate, which in turn helps the code continue with constant step size."

VODE with parameter $MF = 10$ is an implementation of the variable-coefficient Adams method in Nordsieck form (6.30). It is due to Brown, Byrne & Hindmarsh (1989) and supersedes the older code EPISODE of Byrne & Hindmarsh (1975). The authors recommend their code "for problems with widely different active time scales". We used the version of August 31, 1992. It can be obtained by sending an electronic mail to "netlib@research.att.com" with the message

send vode.f from ode	to obtain double precision VODE,
send sode.f from ode	to obtain single precision VODE.

The code VODE differs in several respects from DEABM. The nonlinear equation (first component of (6.30)) is solved by fixed-point iteration until convergence. No final f -evaluation is performed. This method can thus be interpreted as a $P(EC)^M$ -method, where M , the number of iterations, may be different from step to step. E.g., in the example of Fig. 7.2 ($Tol = 10^{-8}$) only 930 function evaluations are needed for 535 steps (519 accepted and 16 rejected). This shows that for many steps one iteration is sufficient. The order selection in VODE is based on maximizing the step size among $h_{\text{opt}}^{(k)}$, $h_{\text{opt}}^{(k+1)}$, $h_{\text{opt}}^{(k+2)}$. Fig. 7.2 presents the step size and order variation for VODE for the same example as above: compared to DEABM we observe that much lower orders are taken. Further, the order is constant over long intervals. This is reasonable, since a change in the order is not natural for the Nordsieck representation.

LSODE (with parameter $MF = 10$) is another implementation of the Adams methods. This is a successor of the code GEAR (Hindmarsh 1972), which is itself a revised and improved code based on DIFSUB of Gear (1971). We used the version of March 30, 1987. LSODE is based on the Nordsieck representation of the fixed step size Adams formulas. It has the same interface as VODE and can be obtained by sending an electronic mail to "netlib@research.att.com" with the message

send lsode.f from odepack

to obtain the double precision version. Fig. 7.3 shows the step sizes and orders chosen by this code. It behaves similarly to VODE.

Numerical Comparisons

Of the three families of methods, the fixed order Runge-Kutta is the simplest, in several respects the best understood, and the least efficient. (Shampine & Gordon 1975)

It is, of course, interesting to study the numerical performance of the above implementations of the Adams methods:

DEABM — symbol ✕

VODE — symbol ○

LSODE — symbol △

In order to compare the results with those of a typical one-step Runge-Kutta method we include the results of the code

DOP853 — symbol ☆

described in Section II.5.

With all these methods we have computed the numerical solution for the six problems EULR, AREN, LRNZ, PLEI, ROPE, BRUS of Section II.10 using many different tolerances between 10^{-3} and 10^{-14} (the “integer” tolerances 10^{-3} , 10^{-4} , . . . are distinguished by enlarged symbols). Fig. 7.4 gives the number of *function evaluations* plotted against the achieved accuracy in double logarithmic scale. Some general tendencies can be distinguished in the crowds of numerical results. LSODE and DEABM require, for equal obtained accuracy, usually less function evaluations, with DEABM becoming champion for higher precision ($Tol \leq 10^{-6}$).

The situation changes dramatically in favour of the Runge-Kutta code DOP853 if *computing time* is measured instead of function evaluations (see Fig. 7.5; the CPU time is that of a Sun Workstation, SunBlade 100). We observe that for problems with cheap function evaluations (EULR, AREN, LRNZ) the Runge-Kutta code needs much less CPU time than the multistep codes, although more function evaluations are necessary in general. For the problems PLEI and ROPE, where the right hand side is rather expensive to evaluate, the discrepancy is not as large. For the last problem (BRUS) the dimension is very high, but the individual components are not too complicated. In this situation, the CPU time of DOP853 is also significantly less than for the multistep codes; this indicates that their overhead also increases with the dimension of the problem.

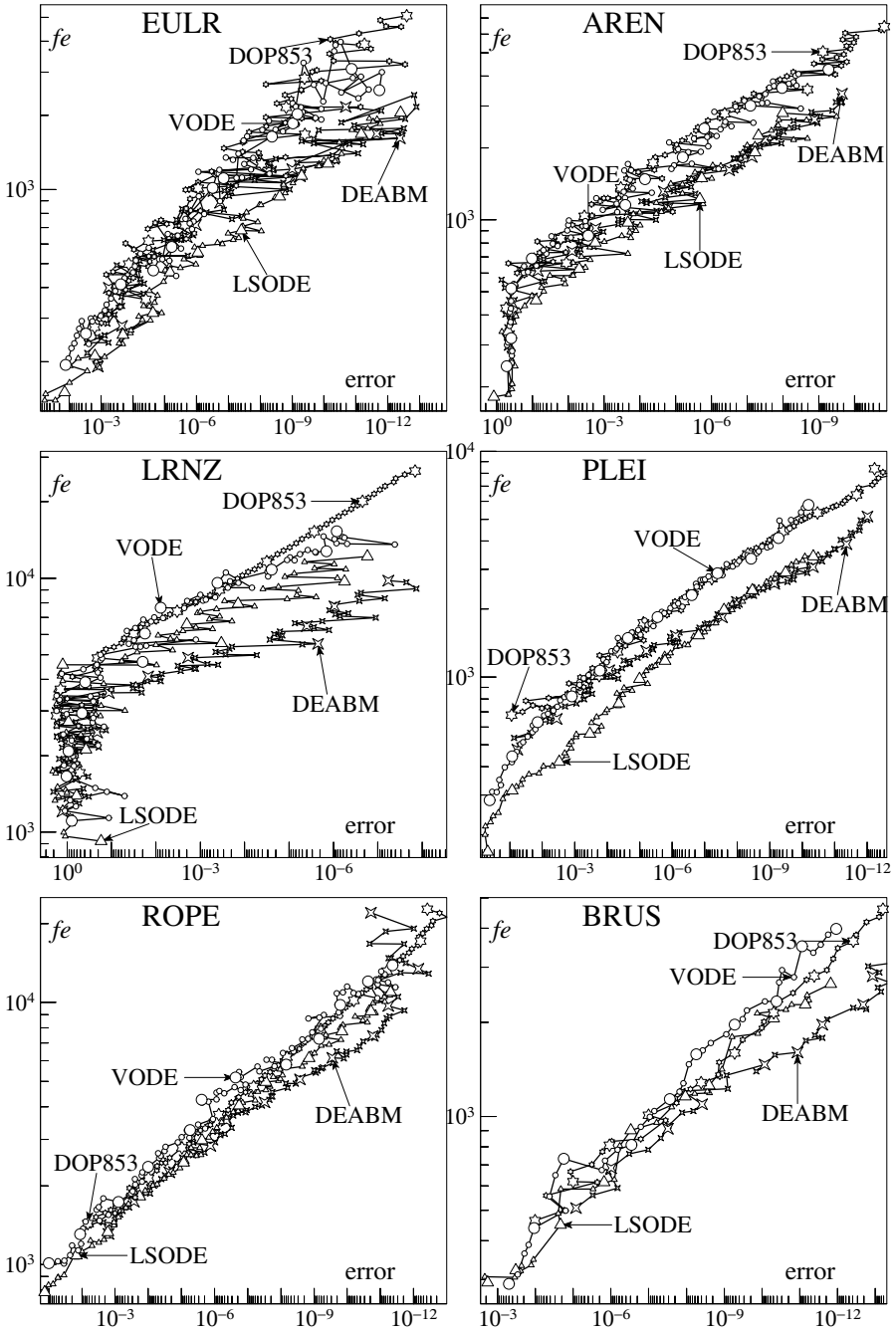


Fig. 7.4. Precision versus function calls for the problems of Section II.10

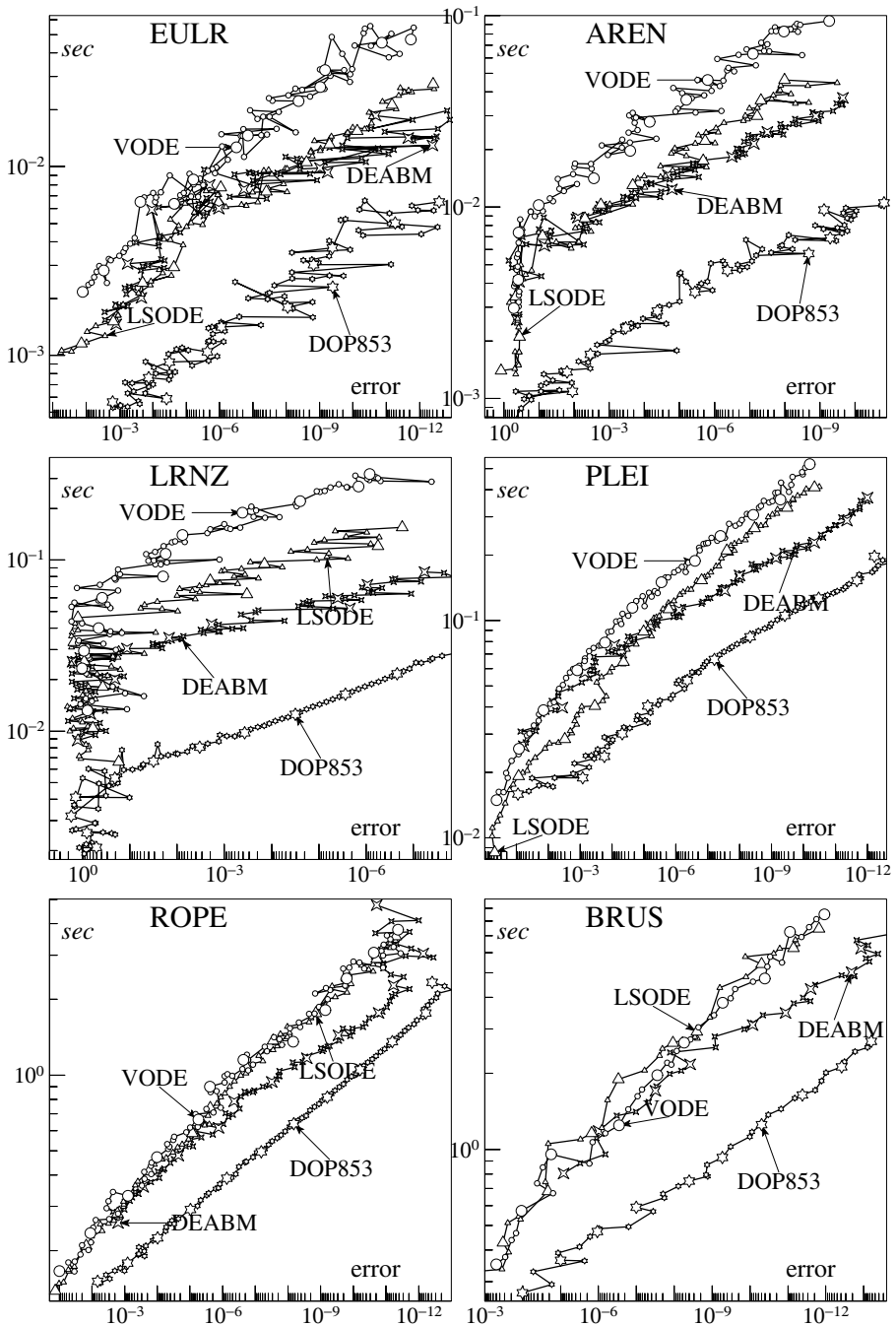


Fig. 7.5. Precision versus computing time for the problems of Section II.10

III.8 General Linear Methods

... methods sufficiently general as to include linear multistep and Runge-Kutta methods as special cases ...

(K. Burrage & J.C. Butcher 1980)

In a remarkably short period (1964-1966) many independent papers appeared which tried to generalize either Runge-Kutta methods in the direction of multistep or multistep methods in the direction of Runge-Kutta. The motivation was either to make the advantages of multistep accessible to Runge-Kutta methods or to “break the Dahlquist barrier” by modifying the multistep formulas. “Generalized multistep methods” were introduced by Gragg and Stetter in (1964), “modified multistep methods” by Butcher (1965a), and in the same year there appeared the work of Gear (1965) on “hybrid methods”. A year later Byrne and Lambert (1966) published their work on “pseudo Runge-Kutta methods”. All these methods fall into the class of “general linear methods” to be discussed in this section.

An example of such a method is the following (Butcher (1965a), order 5)

$$\begin{aligned}\hat{y}_{n+1/2} &= y_{n-1} + \frac{h}{8}(9f_n + 3f_{n-1}) \\ \hat{y}_{n+1} &= \frac{1}{5}(28y_n - 23y_{n-1}) + \frac{h}{5}(32\hat{f}_{n+1/2} - 60f_n - 26f_{n-1}) \\ y_{n+1} &= \frac{1}{31}(32y_n - y_{n-1}) + \frac{h}{93}(64\hat{f}_{n+1/2} + 15\hat{f}_{n+1} + 12f_n - f_{n-1}).\end{aligned}\tag{8.1}$$

We now have the choice of developing a theory of “generalized” multistep methods or of developing a theory of “generalized” Runge-Kutta methods. After having seen in Section III.4 that the convergence theory becomes much nicer when multistep methods are interpreted as one-step methods in higher dimension, we choose the second possibility: since formula (8.1) uses y_n and y_{n-1} as previous information, we introduce the vector $u_n = (y_n, y_{n-1})^T$ so that the last line of (8.1) becomes

$$\begin{pmatrix} y_{n+1} \\ y_n \end{pmatrix} = \begin{pmatrix} \frac{32}{31} & -\frac{1}{31} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} y_n \\ y_{n-1} \end{pmatrix} + \begin{pmatrix} \frac{64}{93} & \frac{15}{93} & \frac{12}{93} & -\frac{1}{93} \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} hf(\hat{y}_{n+1/2}) \\ hf(\hat{y}_{n+1}) \\ hf(y_n) \\ hf(y_{n-1}) \end{pmatrix}$$

which, together with lines 1 and 2 of (8.1), is of the form

$$u_{n+1} = Su_n + h\Phi(x_n, u_n, h).\tag{8.2}$$

Properties of such general methods have been investigated by Butcher (1966),

Hairer & Wanner (1973), Skeel (1976), Cooper (1978), Albrecht (1978, 1985) and others. Clearly, nothing prevents us from letting S and Φ be arbitrary, or from allowing also other interpretations of u_n .

A General Integration Procedure

We consider the system

$$y' = f(x, y), \quad y(x_0) = y_0 \quad (8.3)$$

where f satisfies the regularity condition (4.2). Let m be the dimension of the differential equation (8.3), $q \geq m$ be the dimension of the difference equation (8.2) and $x_n = x_0 + nh$ be the subdivision points of an equidistant grid. The methods under consideration consist of three parts:

- i) a *forward step procedure*, i.e., a formula (8.2), where the square matrix S is independent of (8.3).
- ii) a *correct value function* $z(x, h)$, which gives an interpretation of the values u_n ; $z_n = z(x_n, h)$ is to be approximated by u_n , so that the global error is given by $u_n - z_n$. It is assumed that the exact solution $y(x)$ of (8.3) can be recovered from $z(x, h)$.
- iii) a *starting procedure* $\varphi(h)$, which specifies the starting value $u_0 = \varphi(h)$. $\varphi(h)$ approximates $z_0 = z(x_0, h)$.

The discrete problem corresponding to (8.3) is thus given by

$$u_0 = \varphi(h), \quad (8.4a)$$

$$u_{n+1} = Su_n + h\Phi(x_n, u_n, h), \quad n = 0, 1, 2, \dots, \quad (8.4b)$$

which yields the numerical solution u_0, u_1, u_2, \dots . We remark that the increment function $\Phi(x, u, h)$, the starting procedure $\varphi(h)$ and the correct value function $z(x, h)$ depend on the differential equation (8.3), although this is not stated explicitly.

Example 8.1. The most simple cases are *one-step methods*. A characteristic feature of these is that the dimensions of the differential and difference equation are equal (i.e., $m = q$) and that S is the identity matrix. Furthermore, $\varphi(h) = y_0$ and $z(x, h) = y(x)$. They have been investigated in Chapter II.

Example 8.2. We have seen in Section III.4 that linear *multistep methods* also fall into the class (8.4). For k -step methods the dimension of the difference equation is $q = km$ and the forward step procedure is given by formula (4.8). A starting procedure yields the vector $\varphi(h) = (y_{k-1}, \dots, y_1, y_0)^T$ and, finally, the correct value function is given by

$$z(x, h) = (y(x + (k-1)h), \dots, y(x+h), y(x))^T.$$

The most common way of implementing an implicit multistep method is a *predictor-corrector* process (compare (1.11) and Section III.7): an approximation $y_{n+k}^{(0)}$ to y_{n+k} is “predicted” by an explicit multistep method, say

$$\alpha_k^p y_{n+k}^{(0)} + \alpha_{k-1}^p y_{n+k-1} + \dots + \alpha_0^p y_n = h(\beta_{k-1}^p f_{n+k-1} + \dots + \beta_0^p f_n) \quad (8.5;P)$$

and is then “corrected” (usually once or twice) by

$$f_{n+k}^{(l-1)} := f(x_{n+k}, y_{n+k}^{(l-1)}) \quad (8.5;E)$$

$$\alpha_k y_{n+k}^{(l)} + \alpha_{k-1} y_{n+k-1} + \dots + \alpha_0 y_n = h(\beta_k f_{n+k}^{(l-1)} + \beta_{k-1} f_{n+k-1} + \dots + \beta_0 f_n). \quad (8.5;C)$$

If the iteration (8.5) is carried out until convergence, the process is identical to that of Example 8.2. In practice, however, only a fixed number, say M , of iterations are carried out and the method is theoretically no longer a “pure” multistep method. We distinguish two predictor-corrector (PC) methods, depending on whether it ends with a correction (8.5;C) or not. The first algorithm is symbolized as $P(EC)^M$ and the second possibility, where f_{n+k} is once more updated by (8.5;E) for further use in the subsequent steps, as $P(EC)^M E$. We shall now see how these two procedures can be interpreted as methods of type (8.4).

Example 8.2a. $P(EC)^M E$ -methods. The starting procedure and the correct value function are the same as for multistep methods and also $q = km$. Furthermore we have $S = A \otimes I$, where A is given by (4.7) and I is the m -dimensional identity matrix. Observe that S depends only on the corrector-formula and not on the predictor-formula. Here, the increment function is given by

$$\Phi(x, u, h) = (e_1 \otimes I) \psi(x, u, h)$$

with $e_1 = (1, 0, \dots, 0)^T$. For $u = (u^1, \dots, u^k)^T$ with $u^j \in \mathbb{R}^m$ the function $\psi(x, u, h)$ is defined by

$$\begin{aligned} \psi(x, u, h) = & \alpha_k^{-1} \left(\beta_k f(x + kh, y^{(M)}) \right. \\ & \left. + \beta_{k-1} f(x + (k-1)h, u^1) + \dots + \beta_0 f(x, u^k) \right) \end{aligned}$$

where the value $y^{(M)}$ is calculated from

$$\begin{aligned} & \alpha_k^p y^{(0)} + \alpha_{k-1}^p u^1 + \dots + \alpha_0^p u^k \\ & = h \left(\beta_{k-1}^p f(x + (k-1)h, u^1) + \dots + \beta_0^p f(x, u^k) \right) \\ & \alpha_k y^{(l)} + \alpha_{k-1} u^1 + \dots + \alpha_0 u^k \\ & = h \left(\beta_k f(x + kh, y^{(l-1)}) + \beta_{k-1} f(x + (k-1)h, u^1) + \dots + \beta_0 f(x, u^k) \right) \end{aligned}$$

(for $l = 1, \dots, M$).

Example 8.2b. For P(EC)^M-methods, the formulation as a method of type (8.4) becomes more complicated, since the information to be carried over to the next step is determined not only by y_{n+k-1}, \dots, y_n , but also depends on the values hf_{n+k-1}, \dots, hf_n , where $hf_{n+j} = hf(x_{n+j}, y_{n+j}^{(M-1)})$. Therefore the dimension of the difference equation becomes $q = 2km$. A usual starting procedure (as for multistep methods) yields

$$\varphi(h) = \left(y_{k-1}, \dots, y_0, hf(x_{k-1}, y_{k-1}), \dots, hf(x_0, y_0) \right)^T.$$

If we define the correct value function by

$$z(x, h) = \left(y(x + (k-1)h), \dots, y(x), hy'(x + (k-1)h), \dots, hy'(x) \right)^T,$$

the forward step procedure is given by

$$S = \begin{pmatrix} A & B \\ 0 & N \end{pmatrix}, \quad \Phi(x, u, h) = \begin{pmatrix} \beta'_k e_1 \\ e_1 \end{pmatrix} \Psi(x, u, h).$$

Here A is the matrix given by (4.7), $\beta'_j = \beta_j / \alpha_k$ and

$$N = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} \beta'_{k-1} & \dots & \beta'_0 \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix}, \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

For $u = (u^1, \dots, u^k, hv^1, \dots, hv^k)$ the function $\psi(x, u, h) \in \mathbb{R}^q$ is defined by

$$\psi(x, u, h) = f(x + kh, y^{(M-1)})$$

where $y^{(M-1)}$ is given by

$$\begin{aligned} \alpha_k^p y^{(0)} + \alpha_{k-1}^p u^1 + \dots + \alpha_0^p u^k &= h(\beta_{k-1}^p v^1 + \dots + \beta_0^p v^k) \\ \alpha_k y^{(l)} + \alpha_{k-1} u^1 + \dots + \alpha_0 u^k &= h(\beta_k f(x + kh, y^{(l-1)}) + \beta_{k-1} v^1 + \dots + \beta_0 v^k). \end{aligned}$$

Again we observe that S depends only on the corrector-formula.

Example 8.3. *Nordsieck methods* are also of the form (8.4). This follows immediately from the representation (6.8). In this case the correct value function

$$z(x, h) = \left(y(x), hy'(x), \frac{h^2}{2!} y''(x), \dots, \frac{h^k}{k!} y^{(k)}(x) \right)^T$$

is composed not only of values of the exact solution, but also contains their derivatives.

Example 8.4. *Cyclic multistep methods.* Donelson & Hansen (1971) have investigated the possibility of basing a discretization scheme on several different k -step methods which are used cyclically. Let S_j and Φ_j represent the forward step procedure of the j th multistep method; then the numerical solution u_0, u_1, \dots is

defined by

$$\begin{aligned} u_0 &= \varphi(h) \\ u_{n+1} &= S_j u_n + h \Phi_j(x_n, u_n, h) \quad \text{if } n \equiv (j-1) \bmod m. \end{aligned}$$

In order to get a method (8.4) with S independent of the step number, we consider one cycle of the method as one step of a new method

$$\begin{aligned} u_0^* &= \varphi\left(\frac{h^*}{m}\right) \\ u_{n+1}^* &= S u_n^* + h^* \Phi(x_n^*, u_n^*, h^*) \end{aligned} \quad (8.6)$$

with step size $h^* = mh$. Here $x_n^* = x_0 + nh^*$, $S = S_m \dots S_2 S_1$ and Φ has to be chosen suitably. E.g., in the case $m = 2$ we have

$$\begin{aligned} \Phi(x^*, u^*, h^*) &= \frac{1}{2} S_2 \Phi_1\left(x^*, u^*, \frac{h^*}{2}\right) \\ &\quad + \frac{1}{2} \Phi_2\left(x^* + \frac{h^*}{2}, S_1 u^* + \frac{h^*}{2} \Phi_1\left(x^*, u^*, \frac{h^*}{2}\right), \frac{h^*}{2}\right). \end{aligned}$$

It is interesting to note that cyclically used k -step methods can lead to convergent methods of order $2k - 1$ (or even $2k$). The “first Dahlquist barrier” (Theorem 3.5) can be broken in this way. For more details see Stetter (1973), Albrecht (1979) and Exercise 2.

Example 8.5. General linear methods.

Following the advice of Aristotle ... (the original Greek can be found in Butcher’s paper) ... we look for the greatest good as a mean between extremes. (J.C. Butcher 1985a)

Introduced by Burrage & Butcher (1980), these methods are general enough to include all previous examples as special cases, but at the same time the increment function is given explicitly in terms of the differential equation and several free parameters. They are defined by

$$v_i^{(n)} = \sum_{j=1}^k \tilde{a}_{ij} u_j^{(n)} + h \sum_{j=1}^s \tilde{b}_{ij} f(x_n + c_j h, v_j^{(n)}) \quad i = 1, \dots, s, \quad (8.7a)$$

$$u_i^{(n+1)} = \sum_{j=1}^k a_{ij} u_j^{(n)} + h \sum_{j=1}^s b_{ij} f(x_n + c_j h, v_j^{(n)}) \quad i = 1, \dots, k. \quad (8.7b)$$

The stages $v_i^{(n)}$ ($i = 1, \dots, s$) are the *internal stages* and do not leave the “black box” of the current step. The stages $u_i^{(n)}$ ($i = 1, \dots, k$) are called the *external stages* since they contain all the necessary information from the previous step used in carrying out the current step. The coefficients a_{ij} in (8.7b) form the matrix S of (8.4b). Very often, some internal stages are identical to external ones, as for

example in method (8.1), where

$$v_n = (\hat{y}_{n+1/2}, \hat{y}_{n+1}, y_n, y_{n-1})^T.$$

One-step Runge-Kutta methods are characterized by $k = 1$. At the end of this section we shall discuss the algebraic conditions for general linear methods to be of order p .

Example 8.6. In order to illustrate the fact that the analysis of this section is not only applicable to numerical methods that discretize first order differential equations, we consider the second order initial value problem

$$y'' = g(x, y), \quad y(x_0) = y_0, \quad y'(x_0) = y'_0 \quad (8.8)$$

Replacing $y''(x)$ by a central difference yields

$$y_{n+1} - 2y_n + y_{n-1} = h^2 g(x_n, y_n),$$

and with the additional variables

$$hy'_n = y_{n+1} - y_n$$

this method can be written as

$$\begin{pmatrix} y_{n+1} \\ y'_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} y_n \\ y'_n \end{pmatrix} + h \begin{pmatrix} y'_n \\ g(x_{n+1}, y_n + hy'_n) \end{pmatrix}.$$

It now has the form of a method (8.4) with the correct value function $z(x, h) = (y(x), (y(x+h) - y(x))/h)^T$. Here $y(x)$ denotes the exact solution of (8.8).

Clearly, all Nyström methods (Section II.14) fit into this framework, as do multistep methods for second order differential equations. They will be investigated in more detail in Section III.10.

Example 8.7. *Multi-step multi-stage multi-derivative* methods seem to be the most general class of explicitly given linear methods and generalize the methods of Section II.13. In the notation of that section, we can write

$$v_i^{(n)} = \sum_{j=1}^k \tilde{a}_{ij} u_j^{(n)} + \sum_{r=1}^q \frac{h^r}{r!} \sum_{j=1}^s \tilde{b}_{ij}^{(r)} D^r y(x_n + c_j h, v_j^{(n)}) \quad i = 1, \dots, s,$$

$$u_i^{(n+1)} = \sum_{j=1}^k a_{ij} u_j^{(n)} + \sum_{r=1}^q \frac{h^r}{r!} \sum_{j=1}^s b_{ij}^{(r)} D^r y(x_n + c_j h, v_j^{(n)}) \quad i = 1, \dots, k.$$

Such methods have been studied in Hairer & Wanner (1973).

Stability and Order

The following study of stability, order and convergence follows mainly the lines of Skeel (1976). Stability of a numerical scheme just requires that for $h \rightarrow 0$ the numerical solution remain bounded. This motivates the following definition.

Definition 8.8. Method (8.4) is called *stable* if $\|S^n\|$ is uniformly bounded for all $n \geq 0$.

The local error of method (8.4) is defined in exactly the same way as for one-step methods (Section II.3) and multistep methods (Section III.2).

Definition 8.9. Let $z(x, h)$ be the correct value function for the method (8.4) and let $z_n = z(x_n, h)$. The *local error* is then given by (see Fig. 8.1)

$$\begin{aligned} d_0 &= z_0 - \varphi(h) \\ d_{n+1} &= z_{n+1} - Sz_n - h\Phi(x_n, z_n, h), \quad n = 0, 1, \dots \end{aligned} \quad (8.9)$$

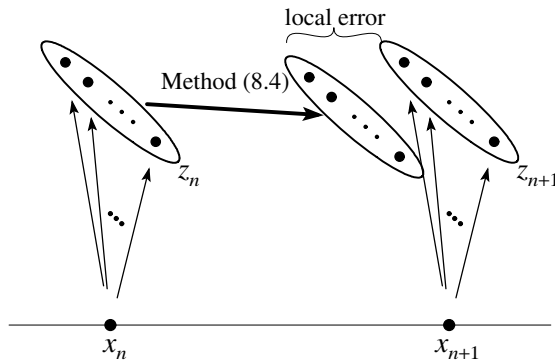


Fig. 8.1. Illustration of the local error

The definition of order is not as straightforward. The requirement that the local error be $\mathcal{O}(h^{p+1})$ (cf. one-step and multistep methods) will turn out to be sufficient but in general not necessary for convergence of order p . For an appropriate definition we need the *spectral decomposition* of the matrix S .

First observe that, whenever the local error (8.9) tends to zero for $h \rightarrow 0$ ($nh = x - x_0$ fixed), we get

$$0 = z(x, 0) - Sz(x, 0), \quad (8.10)$$

so that 1 is an eigenvalue of S and $z(x, 0)$ a corresponding eigenvector. Furthermore, by stability, no eigenvalue of S can lie outside the unit disc and the eigenvalues of modulus one can not give rise to Jordan chains. Denoting the eigenvalues of modulus one by $\zeta_1 (= 1), \zeta_2, \dots, \zeta_l$, the Jordan canonical form of S (see

(I.12.14)) is therefore the block diagonal matrix

$$S = T \operatorname{diag} \left\{ \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}, \begin{pmatrix} \zeta_2 & & \\ & \ddots & \\ & & \zeta_2 \end{pmatrix}, \dots, \begin{pmatrix} \zeta_l & & \\ & \ddots & \\ & & \zeta_l \end{pmatrix}, \tilde{J} \right\} T^{-1}.$$

If we decompose this matrix into the terms which correspond to the single eigenvalues we obtain

$$S = E + \zeta_2 E_2 + \dots + \zeta_l E_l + \tilde{E} \quad (8.11)$$

where

$$E = T \operatorname{diag} \left\{ I, 0, 0, \dots \right\} T^{-1}, \quad (8.12)$$

$$E_2 = T \operatorname{diag} \left\{ 0, I, 0, \dots \right\} T^{-1}, \dots, E_l = T \operatorname{diag} \left\{ 0, \dots, 0, I, 0 \right\} T^{-1},$$

$$\tilde{E} = T \operatorname{diag} \left\{ 0, 0, 0, \dots, \tilde{J} \right\} T^{-1}.$$

We are now prepared to give

Definition 8.10. The method (8.4) is of *order p (consistent of order p)*, if for all problems (8.3) with p times continuously differentiable f , the local error satisfies

$$d_0 = \mathcal{O}(h^p) \\ E(d_0 + d_1 + \dots + d_n) + d_{n+1} = \mathcal{O}(h^p) \quad \text{for } 0 \leq nh \leq \text{Const.} \quad (8.13)$$

Remark. This property is called *quasi-consistency of order p* by Skeel (1976).

If the right-hand side of the differential equation (8.3) is p -times continuously differentiable then, in general, $\varphi(h)$, $\Phi(x, u, h)$ and $z(x, h)$ are also smooth, so that the local error (8.9) can be expanded into a Taylor series in h :

$$d_0 = \gamma_0 + \gamma_1 h + \dots + \gamma_{p-1} h^{p-1} + \mathcal{O}(h^p) \\ d_{n+1} = \delta_0(x_n) + \delta_1(x_n) h + \dots + \delta_p(x_n) h^p + \mathcal{O}(h^{p+1}). \quad (8.14)$$

The function $\delta_j(x)$ is then $(p-j+1)$ -times continuously differentiable. The following lemma gives a more practical characterization of the order of the methods (8.4).

Lemma 8.11. Assume that the local error of method (8.4) satisfies (8.14) with continuous $\delta_j(x)$. The method is then of order p , if and only if

$$d_n = \mathcal{O}(h^p) \quad \text{for } 0 \leq nh \leq \text{Const}, \quad \text{and} \quad E\delta_p(x) = 0. \quad (8.15)$$

Proof. The condition (8.15) is equivalent to

$$d_n = \mathcal{O}(h^p), \quad E d_{n+1} = \mathcal{O}(h^{p+1}) \quad \text{for } 0 \leq nh \leq \text{Const}, \quad (8.16)$$

which is clearly sufficient for order p . We now show that (8.15) is also necessary. Since $E^2 = E$ (see (8.12)) order p implies

$$d_n = \mathcal{O}(h^p), \quad E(d_1 + \dots + d_n) = \mathcal{O}(h^p) \quad \text{for } 0 \leq nh \leq \text{Const.} \quad (8.17)$$

This is best seen by multiplying (8.13) by E . Consider now pairs (n, h) such that $nh = x - x_0$ for some fixed x . We insert (8.14) (observe that $d_n = \mathcal{O}(h^p)$) into $E(d_1 + \dots + d_n)$ and approximate the resulting sum by the corresponding Riemann integral

$$E(d_1 + \dots + d_n) = h^p E \sum_{j=1}^n \delta_p(x_{j-1}) + \mathcal{O}(h^p) = h^{p-1} E \int_{x_0}^x \delta_p(s) ds + \mathcal{O}(h^p).$$

It follows from (8.17) that $E \int_{x_0}^x \delta_p(s) ds = 0$ and by differentiation that $E\delta_p(x) = 0$. \square

Convergence

In addition to the numerical solution given by (8.4) we consider a perturbed numerical solution (\hat{u}_n) defined by

$$\begin{aligned} \hat{u}_0 &= \varphi(h) + r_0 \\ \hat{u}_{n+1} &= S\hat{u}_n + h\Phi(x_n, \hat{u}_n, h) + r_{n+1}, \quad n = 0, 1, \dots, N-1 \end{aligned} \quad (8.18)$$

for some perturbation $R = (r_0, r_1, \dots, r_N)$. For example, the exact solution $z_n = z(x_n, h)$ can be interpreted as a perturbed solution, where the perturbation is just the local error. The following lemma gives the best possible qualitative bound on the difference $u_n - \hat{u}_n$ in terms of the perturbation R . We have to assume that the increment function $\Phi(x, u, h)$ satisfies a Lipschitz condition with respect to u (on a compact neighbourhood of the solution). This is the case for all reasonable methods.

Lemma 8.12. *Let the method (8.4) be stable and assume the sequences (u_n) and (\hat{u}_n) be given by (8.4) and (8.18), respectively. Then there exist positive constants c and C such that for any perturbation R and for $hN \leq \text{Const}$*

$$c\|R\|_S \leq \max_{0 \leq n \leq N} \|u_n - \hat{u}_n\| \leq C\|R\|_S$$

with

$$\|R\|_S = \max_{0 \leq n \leq N} \left\| \sum_{j=0}^n S^{n-j} r_j \right\|.$$

Remark. $\|R\|_S$ is a norm on $\mathbb{R}^{(N+1)q}$. Its positivity is seen as follows: if $\|R\|_S = 0$ then for $n = 0, 1, 2, \dots$ one obtains $r_0 = 0, r_1 = 0, \dots$ recursively.

Proof. Set $\Delta u_n = \hat{u}_n - u_n$ and $\Delta \Phi_n = \Phi(x_n, \hat{u}_n, h) - \Phi(x_n, u_n, h)$. Then we have

$$\Delta u_{n+1} = S\Delta u_n + h\Delta \Phi_n + r_{n+1}. \quad (8.19)$$

By assumption there exists a constant L such that $\|\Delta \Phi_n\| \leq L\|\Delta u_n\|$. Solving the difference equation (8.19) gives $\Delta u_0 = r_0$ and

$$\Delta u_{n+1} = \sum_{j=0}^n S^{n-j} h \Delta \Phi_j + \sum_{j=0}^{n+1} S^{n+1-j} r_j. \quad (8.20)$$

By stability there exists a constant B such that

$$\|S^n\|L \leq B \quad \text{for all } n \geq 0. \quad (8.21)$$

Thus (8.20) becomes

$$\|\Delta u_{n+1}\| \leq hB \sum_{j=0}^n \|\Delta u_j\| + \|R\|_S.$$

By induction on n it follows that

$$\|\Delta u_n\| \leq (1 + hB)^n \|R\|_S \leq \exp(\text{Const} \cdot B) \cdot \|R\|_S,$$

which proves the second inequality in the lemma. From (8.20) and (8.21)

$$\left\| \sum_{j=0}^n S^{n-j} r_j \right\| \leq (1 + nhB) \max_{0 \leq n \leq N} \|\Delta u_n\|,$$

and we thus obtain for $Nh \leq \text{Const}$

$$\|R\|_S \leq (1 + \text{Const} \cdot B) \cdot \max_{0 \leq n \leq N} \|\hat{u}_n - u_n\|. \quad \square$$

Remark. Two-sided error bounds, such as in Lemma 8.12, were first studied, in the case of multi-step methods, by Spijker (1971). This theory has become prominent through the treatment of Stetter (1973, pp. 81-84). Extensions to general linear methods are due to Skeel (1976) and Albrecht (1978).

Using the lemma above we can prove

Theorem 8.13. *Consider a stable method (8.4) and assume that the local error satisfies (8.14) with $\delta_p(x)$ continuously differentiable. The method is then convergent of order p , i.e., the global error $u_n - z_n$ satisfies*

$$u_n - z_n = \mathcal{O}(h^p) \quad \text{for } 0 \leq nh \leq \text{Const},$$

if and only if it is consistent of order p .

Proof. The identity

$$E(d_0 + \dots + d_n) + d_{n+1} = \sum_{j=0}^{n+1} S^{n+1-j} d_j - (S - E) \sum_{j=0}^n S^{n-j} d_j,$$

which is a consequence of $ES = E$ (see (8.11) and (8.12)), implies that for $n \leq N - 1$ and $D = (d_0, \dots, d_N)$,

$$\|E(d_0 + \dots + d_n) + d_{n+1}\| \leq (1 + \|S - E\|) \cdot \|D\|_S. \quad (8.22)$$

The lower bound of Lemma 8.12, with r_n and \hat{u}_n replaced by d_n and z_n respectively, yields the “only if” part of the theorem.

For the “if” part we use the upper bound of Lemma 8.12. We have to show that consistency of order p implies

$$\max_{0 \leq n \leq N} \left\| \sum_{j=0}^n S^{n-j} d_j \right\| = \mathcal{O}(h^p). \quad (8.23)$$

By (8.11) and (8.12) we have

$$S^{n-j} = E + \zeta_2^{n-j} E_2 + \dots + \zeta_l^{n-j} E_l + \tilde{E}^{n-j}.$$

This identity together with Lemma 8.11 implies

$$\begin{aligned} \sum_{j=0}^n S^{n-j} d_j &= h^p E_2 \sum_{j=1}^n \zeta_2^{n-j} \delta_p(x_{j-1}) + \dots \\ &\quad + h^p E_l \sum_{j=1}^n \zeta_l^{n-j} \delta_p(x_{j-1}) + \sum_{j=0}^n \tilde{E}^{n-j} d_j + \mathcal{O}(h^p). \end{aligned}$$

The last term in this expression is $\mathcal{O}(h^p)$ since in a suitable norm $\|\tilde{E}\| < 1$ and therefore

$$\left\| \sum_{j=0}^n \tilde{E}^{n-j} d_j \right\| \leq \sum_{j=0}^n \|\tilde{E}\|^{n-j} \|d_j\| \leq \frac{1}{1 - \|\tilde{E}\|} \cdot \max_{0 \leq n \leq N} \|d_n\|.$$

For the rest we use partial summation (Abel 1826)

$$\sum_{j=1}^n \zeta^{n-j} \delta(x_{j-1}) = \frac{1 - \zeta^n}{1 - \zeta} \cdot \delta(x_0) + \sum_{j=1}^n \frac{1 - \zeta^{n-j}}{1 - \zeta} \cdot (\delta(x_j) - \delta(x_{j-1})) = \mathcal{O}(1),$$

whenever $|\zeta| = 1$, $\zeta \neq 1$ and δ is of bounded variation. \square

Order Conditions for General Linear Methods

For the construction of a p th order general linear method (8.7) the conditions (8.15) are still not very practical. One would like to have instead algebraic conditions in the free parameters, as is the case for Runge-Kutta methods. We shall demonstrate how this can be achieved using the theory of B-series of Section II.12 (see also Burrage & Moss 1980). In order to avoid tensor products we assume in what follows that the differential equation under consideration is a scalar one. All results, however, are also valid for systems. We further assume the differential equation to be autonomous, so that the theory of Section II.12 is directly applicable. This will be justified in Remark 8.17 below.

Suppose now that the components of the correct value function $z(x, h) = (z_1(x, h), \dots, z_k(x, h))^T$ possess an expansion as a B-series

$$z_i(x, h) = B(\mathbf{z}_i, y(x))$$

so that with $\mathbf{z}(t) = (\mathbf{z}_1(t), \dots, \mathbf{z}_k(t))^T$,

$$z(x, h) = \mathbf{z}(\emptyset)y(x) + h\mathbf{z}(\tau)f(y(x)) + \dots \quad (8.24)$$

Before deriving the order conditions we observe that (8.7a) makes sense only if $v_j^{(n)} \rightarrow y(x_n)$ for $h \rightarrow 0$. Otherwise $f(v_j^{(n)})$ need not be defined. Since $u_j^{(n)}$ is an approximation of $z_j(x_n, h)$, this leads to the condition $\sum \tilde{a}_{ij}\mathbf{z}_j(\emptyset) = \mathbb{1}$. This together with (8.10) are the so-called *preconsistency conditions*:

$$A\mathbf{z}(\emptyset) = \mathbf{z}(\emptyset), \quad \tilde{A}\mathbf{z}(\emptyset) = \mathbb{1}. \quad (8.25)$$

A and \tilde{A} are the matrices with entries a_{ij} and \tilde{a}_{ij} , respectively, and $\mathbb{1}$ is the column vector $(1, \dots, 1)^T$. Recall that the local error (8.9) for the general linear method (8.7) is given by

$$d_i^{(n+1)} = z_i(x_n + h, h) - \sum_{j=1}^k a_{ij}z_j(x_n, h) - \sum_{j=1}^s b_{ij}hf(v_j) \quad (8.26a)$$

where

$$v_i = \sum_{j=1}^k \tilde{a}_{ij}z_j(x_n, h) + \sum_{j=1}^s \tilde{b}_{ij}hf(v_j). \quad (8.26b)$$

For the derivation of the order conditions we write v_i and $d_i^{(n+1)}$ as B-series

$$v_i = B(\mathbf{v}_i, y(x_n)), \quad d_i^{(n+1)} = B(\mathbf{d}_i, y(x_n)).$$

By the composition theorem for B-series and by formula (12.10) of Section II.12 we have

$$z_i(x_n + h, h) = B(\mathbf{z}_i, y(x_n + h)) = B(\mathbf{z}_i, B(\mathbf{p}, y(x_n))) = B(\mathbf{p}\mathbf{z}_i, y(x_n)).$$

Inserting all these series into (8.26) and comparing the coefficients we arrive at

$$\begin{aligned} \mathbf{d}_i(t) &= (\mathbf{p}\mathbf{z}_i)(t) - \sum_{j=1}^k a_{ij} \mathbf{z}_j(t) - \sum_{j=1}^s b_{ij} \mathbf{v}'_j(t) \\ \mathbf{v}_i(t) &= \sum_{j=1}^k \tilde{a}_{ij} \mathbf{z}_j(t) + \sum_{j=1}^s \tilde{b}_{ij} \mathbf{v}'_j(t). \end{aligned} \quad (8.27)$$

An application of Lemma 8.11 now yields

Theorem 8.14. *Let $\mathbf{d}(t) = (\mathbf{d}_1(t), \dots, \mathbf{d}_k(t))^T$ with $\mathbf{d}_i(t)$ be given by (8.27). The general linear method (8.7) is of order p , iff*

$$\begin{aligned} \mathbf{d}(t) &= 0 & \text{for } t \in T, \varrho(t) \leq p-1, \\ E\mathbf{d}(t) &= 0 & \text{for } t \in T, \varrho(t) = p, \end{aligned} \quad (8.28)$$

where the matrix E is defined in (8.12). \square

Corollary 8.15. *Sufficient conditions for the general linear method to be of order p are*

$$\mathbf{d}(t) = 0 \quad \text{for } t \in T, \varrho(t) \leq p. \quad (8.29)$$

\square

Remark 8.16. The expression $(\mathbf{p}\mathbf{z}_i)(t)$ in (8.27) can be computed using formula (12.8) of Section II.12. Since $\mathbf{p}(t) = 1$ for all trees t , we have

$$(\mathbf{p}\mathbf{z}_i)(t) = \sum_{j=0}^{\varrho(t)} \binom{\varrho(t)}{j} \frac{1}{\alpha(t)} \sum_{\text{all labellings}} \mathbf{z}_i(s_j(t)). \quad (8.30)$$

This rather complicated formula simplifies considerably if we assume that the coefficients $\mathbf{z}_i(t)$ of the correct value function depend only on the order of t , i.e., that

$$\mathbf{z}_i(t) = \mathbf{z}_i(u) \quad \text{whenever } \varrho(t) = \varrho(u). \quad (8.31)$$

In this case formula (8.30) becomes

$$(\mathbf{p}\mathbf{z}_i)(t) = \sum_{j=0}^{\varrho(t)} \binom{\varrho(t)}{j} \mathbf{z}_i(\tau^j). \quad (8.32)$$

Here τ^j represents any tree of order j , e.g.,

$$\tau^j = [\underbrace{\tau, \dots, \tau}_{j-1}], \quad \tau^1 = \tau, \quad \tau^0 = \emptyset. \quad (8.33)$$

Usually the components of $z(x, h)$ are composed of

$$y(x), y(x + jh), hy'(x), h^2y''(x), \dots,$$

in which case assumption (8.31) is satisfied.

Remark 8.17. Non-autonomous systems. For the differential equation $x' = 1$, formula (8.7a) becomes

$$v_n = \tilde{A}u_n + h\tilde{B}\mathbb{I}.$$

Assuming that $x' = 1$ is integrated exactly, i.e., $u_n = z(\emptyset)x_n + hz(\tau)$ we obtain $v_n = x_n\mathbb{I} + hc$, where $c = (c_1, \dots, c_s)^T$ is given by

$$c = \tilde{A}z(\tau) + \tilde{B}e. \quad (8.34)$$

This definition of the c_i implies that the numerical results for $y' = f(x, y)$ and for the augmented autonomous differential equation are the same and the above results are also valid in the general case.

Table 8.1 presents the order conditions up to order 3 in addition to the preconsistency conditions (8.25). We assume that (8.31) is satisfied and that c is given by (8.34). Furthermore, c^j denotes the vector $(c_1^j, \dots, c_s^j)^T$.

Table 8.1. Order conditions for general linear methods

t	$\varrho(t)$	order condition
τ	1	$Az(\tau) + B\mathbb{I} = z(\tau) + z(\emptyset)$
τ^2	2	$Az(\tau^2) + 2Bc = z(\tau^2) + 2z(\tau) + z(\emptyset)$
τ^3	3	$Az(\tau^3) + 3Bc^2 = z(\tau^3) + 3z(\tau^2) + 3z(\tau) + z(\emptyset)$
$[\tau^2]$	3	$Az(\tau^3) + 3Bv(\tau^2) = z(\tau^3) + 3z(\tau^2) + 3z(\tau) + z(\emptyset)$ with $v(\tau^2) = \tilde{A}z(\tau^2) + 2\tilde{B}c$

Construction of General Linear Methods

Let us demonstrate on an example how low order methods can be constructed: we set $k = s = 2$ and fix the correct value function as

$$z(x, h) = (y(x), y(x - h))^T.$$

This choice satisfies (8.24) and (8.31) with

$$z(\emptyset) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad z(\tau) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \quad z(\tau^2) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \dots$$

Since the second component of $z(x+h, h)$ is equal to the first component of $z(x, h)$, it is natural to look for methods with

$$A = \begin{pmatrix} a_{11} & a_{12} \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} b_{11} & b_{12} \\ 0 & 0 \end{pmatrix}.$$

We further impose

$$\tilde{B} = \begin{pmatrix} 0 & 0 \\ \tilde{b}_{21} & 0 \end{pmatrix}$$

so that the resulting method is explicit.

The preconsistency condition (8.25), formula (8.34) and the order conditions of Table 8.1 yield the following equations to be solved:

$$a_{11} + a_{12} = 1 \quad (8.35a)$$

$$\tilde{a}_{11} + \tilde{a}_{12} = 1, \quad \tilde{a}_{21} + \tilde{a}_{22} = 1 \quad (8.35b)$$

$$c_1 = -\tilde{a}_{12}, \quad c_2 = \tilde{b}_{21} - \tilde{a}_{22} \quad (8.35c)$$

$$-a_{12} + b_{11} + b_{12} = 1 \quad (8.35d)$$

$$a_{12} + 2(b_{11}c_1 + b_{12}c_2) = 1 \quad (8.35e)$$

$$-a_{12} + 3(b_{11}c_1^2 + b_{12}c_2^2) = 1 \quad (8.35f)$$

$$-a_{12} + 3(b_{11}\tilde{a}_{12} + b_{12}(\tilde{a}_{22} + 2\tilde{b}_{21}c_1)) = 1. \quad (8.35g)$$

These are 9 equations in 11 unknowns. Letting c_1 and c_2 be free parameters, we obtain the solution in the following way: compute a_{12} , b_{11} and b_{12} from the linear system (8.35d,e,f), then \tilde{a}_{12} , \tilde{a}_{22} and \tilde{b}_{21} from (8.35c,g) and finally a_{11} , \tilde{a}_{11} and \tilde{a}_{21} from (8.35a,b). A particular solution for $c_1 = 1/2$, $c_2 = -2/5$ is:

$$A = \begin{pmatrix} 16/11 & -5/11 \\ 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 104/99 & -50/99 \\ 0 & 0 \end{pmatrix}, \quad (8.36)$$

$$\tilde{A} = \begin{pmatrix} 3/2 & -1/2 \\ 3/2 & -1/2 \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} 0 & 0 \\ -9/10 & 0 \end{pmatrix}.$$

This method, which represents a stable explicit 2-step, 2-stage method of order 3, is due to Butcher (1984).

The construction of higher order methods soon becomes very complicated, and the use of *simplifying assumptions* will be very helpful:

Theorem 8.18 (Burrage & Moss 1980). *Assume that the correct value function satisfies (8.31). The simplifying assumptions*

$$\tilde{A}z(\tau^j) + j\tilde{B}c^{j-1} = c^j \quad j = 1, \dots, p-1 \quad (8.37)$$

together with the preconsistency relations (8.25) and the order conditions for the “bushy trees”

$$d(\tau^j) = 0 \quad j = 1, \dots, p$$

imply that the method (8.7) is of order p .

Proof. An induction argument based on (8.27) implies that

$$\mathbf{v}(t) = \mathbf{v}(\tau^j) \quad \text{for } \varrho(t) = j, \quad j = 1, \dots, p-1$$

and consequently also that

$$\mathbf{d}(t) = \mathbf{d}(\tau^j) \quad \text{for } \varrho(t) = j, \quad j = 1, \dots, p. \quad \square$$

The simplifying assumptions (8.37) allow an interesting interpretation: they are equivalent to the fact that the internal stages $v_1^{(n)}$ approximate the exact solution at $x_n + c_i h$ up to order $p-1$, i.e., that

$$v_i^{(n)} - y(x_n + c_i h) = \mathcal{O}(h^p).$$

In the case of Runge-Kutta methods (8.37) reduces to the conditions $C(p-1)$ of Section II.7.

For further examples of general linear methods satisfying (8.37) we refer to Burrage & Moss (1980) and Butcher (1981). See also Burrage (1985) and Butcher (1985a).

Exercises

1. Consider the composition of (cf. Example 8.5)
 - a) explicit and implicit Euler method;
 - b) implicit and explicit Euler method.
 To which methods are they equivalent? What is the order of the composite methods?
2. a) Suppose that each of the m multistep methods (ϱ_i, σ_i) $i = 1, \dots, m$ is of order p . Prove that the corresponding cyclic method is of order at least p .
 b) Construct a stable, 2-cyclic, 3-step linear multistep method of order 5: find first a one-parameter family of linear 3-step methods of order 5 (which are necessarily unstable).

Result.

$$\begin{aligned} \varrho_c(\zeta) &= c\zeta^3 + \left(\frac{19}{30} - c\right)\zeta^2 - \left(\frac{8}{30} + c\right)\zeta + \left(c - \frac{11}{30}\right) \\ \sigma_c(\zeta) &= \left(\frac{1}{9} - \frac{c}{3}\right)\zeta^3 + \left(c + \frac{8}{30}\right)\zeta^2 + \left(\frac{19}{30} - c\right)\zeta + \left(\frac{c}{3} - \frac{1}{90}\right). \end{aligned}$$

Then determine c_1 and c_2 , such that the eigenvalues of the matrix S for the composite method become 1, 0, 0.

3. Prove that the composition of two different general linear methods (with the same correct value function) again gives a general linear method. As a consequence, the cyclic methods of Example 8.4 are general linear methods.

4. Suppose that all eigenvalues of S (except $\zeta_1 = 1$) lie inside the unit circle. Then

$$\|R\|_E = \max_{0 \leq n \leq N} \left\| r_n + E \sum_{j=0}^{n-1} r_j \right\|$$

is a minimal stability functional.

5. Verify for linear multistep methods that the consistency conditions (2.6) are equivalent to consistency of order 1 in the sense of Lemma 8.11.
6. Write method (8.1) as general linear method (8.7) and determine its order (answer: $p = 5$).
7. Interpret the method of Cairra, Costabile & Costabile (1990)

$$k_i^n = hf \left(x_n + c_i h, y_n + \sum_{j=1}^s \bar{a}_{ij} k_j^{n-1} + \sum_{j=1}^{i-1} a_{ij} k_j^n \right)$$

$$y_{n+1} = y_n + \sum_{i=1}^s b_i k_i^n$$

as general linear method. Show that, if

$$\|k_i^{-1} - hy'(x_0 + (c_i - 1)h)\| \leq C \cdot h^p,$$

$$\sum_{i=1}^s b_i c_i^{q-1} = \frac{1}{q}, \quad q = 1, \dots, p,$$

$$\sum_{j=1}^s \bar{a}_{ij} (c_j - 1)^{q-1} + \sum_{j=1}^{i-1} a_{ij} c_j^{q-1} = \frac{c_i^q}{q}, \quad q = 1, \dots, p-1,$$

then the method is of order at least p . Find parallels of these conditions with those of Theorem 8.18.

8. Jackiewicz & Zennaro (1992) propose the following two-step Runge-Kutta method

$$Y_i^{n-1} = y_{n-1} + h_{n-1} \sum_{j=1}^{i-1} a_{ij} f(Y_j^{n-1}), \quad Y_i^n = y_n + h_{n-1} \xi \sum_{j=1}^{i-1} a_{ij} f(Y_j^n),$$

$$y_{n+1} = y_n + h_{n-1} \sum_{i=1}^s v_i f(Y_i^{n-1}) + h_{n-1} \xi \sum_{i=1}^s w_i f(Y_i^n), \quad (8.38)$$

where $\xi = h_n/h_{n-1}$. The coefficients v_i, w_i may depend on ξ , but the a_{ij} do not. Hence, this method requires s function evaluations per step.

- a) Show that the order of method (8.38) is p (according to Definition 8.10) if

and only if for all trees t with $1 \leq \varrho(t) \leq p$

$$\xi^{\varrho(t)} = \sum_{i=1}^s v_i (\mathbf{y}^{-1} \mathbf{g}'_i)(t) + \xi^{\varrho(t)} \sum_{i=1}^s w_i \mathbf{g}'_i(t), \quad (8.39)$$

where, as for Runge-Kutta methods, $\mathbf{g}_i(t) = \sum_{j=1}^{i-1} a_{ij} \mathbf{g}'_j(t)$. The coefficients $\mathbf{y}^{-1}(t) = (-1)^{\varrho(t)}$ are those of $y(x_n - h) = B(\mathbf{y}^{-1}, y(x_n))$.

b) Under the assumption

$$v_i + \xi^p w_i = 0 \quad \text{for } i = 2, \dots, s \quad (8.40)$$

the order conditions (8.39) are equivalent to

$$\xi = \sum_{i=1}^s v_i + \xi \sum_{i=1}^s w_i, \quad (8.41a)$$

$$\xi^r = \sum_{j=1}^{r-1} j \binom{r}{j} (-1)^{r-j} \sum_{i=1}^s v_i c_i^{j-1} + (1 - \xi^{r-p}) r \sum_{i=1}^s v_i c_i^{r-1}, \quad r = 2, \dots, p, \quad (8.41b)$$

$$\sum_{i=1}^s v_i (\mathbf{g}'_i(u) - \varrho(u) c_i^{\varrho(u)-1}) = 0 \quad \text{for } \varrho(u) \leq p-1. \quad (8.41c)$$

c) The conditions (8.41a,b) uniquely define $\sum_i w_i$, $\sum_i v_i c_i^{j-1}$ as functions of $\xi > 0$ (for $j = 1, \dots, p-1$).

d) For each continuous Runge-Kutta method of order $p-1 \geq 2$ there exists a method (8.38) of order p with the same coefficient matrix (a_{ij}) .

Hints. To obtain (8.41c) subtract equation (8.40) from the same equation where t is replaced by the bushy tree of order $\varrho(t)$. Then proceed by induction. The conditions $\sum_i v_i c_i^{j-1} = f_j^p(\xi)$, $j = 1, \dots, p-1$, obtained from (c), together with (8.41c) have the same structure as the order conditions (order $p-1$) of a continuous Runge-Kutta method (Theorem II.6.1).

III.9 Asymptotic Expansion of the Global Error

The asymptotic expansion of the global error of multistep methods was studied in the famous thesis of Gragg (1964). His proof is very technical and can also be found in a modified version in the book of Stetter (1973), pp. 234-245. The existence of asymptotic expansions for general linear methods was conjectured by Skeel (1976). The proof given below (Hairer & Lubich 1984) is based on the ideas of Section II.8.

An Instructive Example

Let us start with an example in order to understand which kind of asymptotic expansion may be expected. We consider the simple differential equation

$$y' = -y, \quad y(0) = 1,$$

take a constant step size h and apply the 3-step BDF-formula (1.22') with one of the following three starting procedures:

$$y_0 = 1, \quad y_1 = \exp(-h), \quad y_2 = \exp(-2h) \quad (\text{exact values}) \quad (9.1a)$$

$$y_0 = 1, \quad y_1 = 1 - h + \frac{h^2}{2} - \frac{h^3}{6}, \quad y_2 = 1 - 2h + 2h^2 - \frac{4h^3}{3}, \quad (9.1b)$$

$$y_0 = 1, \quad y_1 = 1 - h + \frac{h^2}{2}, \quad y_2 = 1 - 2h + 2h^2. \quad (9.1c)$$

The three pictures on the left of Fig. 9.1 (they correspond to the three starting procedures in the same order) show the global error divided by h^3 for the five step sizes $h = 1/5, 1/10, 1/20, 1/40, 1/80$.

For the first two starting procedures we observe uniform convergence to the function $e_3(x) = xe^{-x}/4$ (cf. formula (2.12)), so that

$$y_n - y(x_n) = e_3(x_n)h^3 + \mathcal{O}(h^4), \quad (9.2)$$

valid uniformly for $0 \leq nh \leq \text{Const}$. In the third case we have convergence to $e_3(x) = (9+x)e^{-x}/4$ (Exercise 2), but this time the convergence is no longer uniform. Therefore (9.2) only holds for x_n bounded away from x_0 , i.e., for $0 < \alpha \leq nh \leq \text{Const}$. In the three pictures on the right of Fig. 9.1 the functions

$$(y_n - y(x_n) - e_3(x_n)h^3)/h^4 \quad (9.3)$$

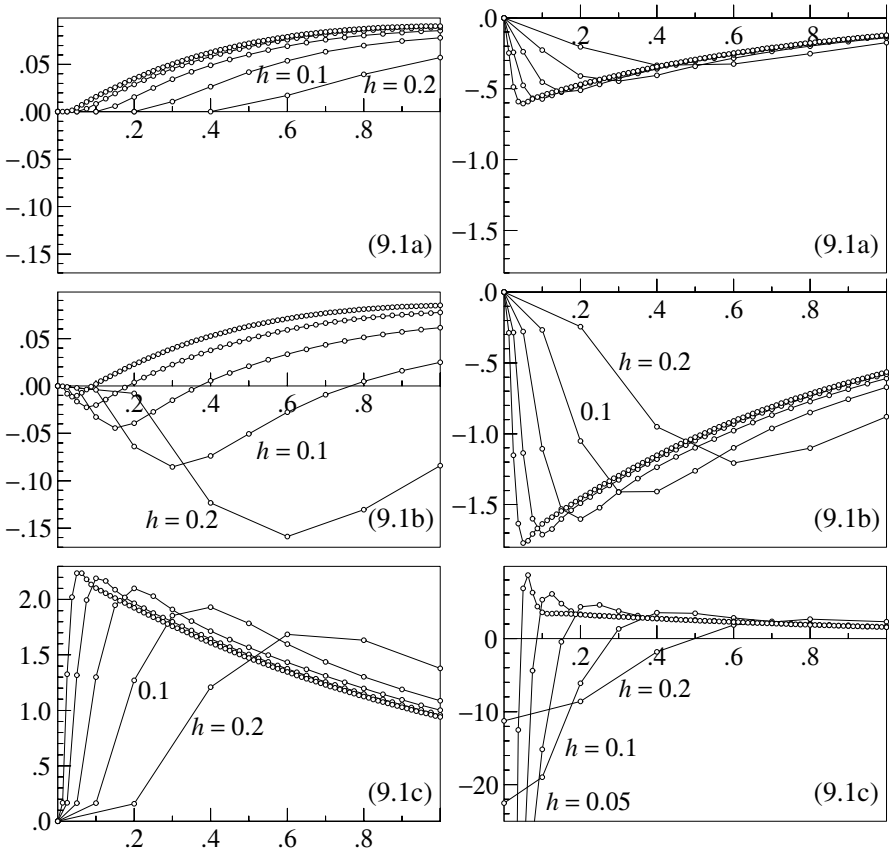


Fig. 9.1. The values $(y_n - y(x_n))/h^3$ (left), $(y_n - y(x_n) - e_3(x_n)h^3)/h^4$ (right) for the 3-step BDF method and for three different starting procedures

are plotted. Convergence to functions $e_4(x)$ is observed in all cases. Clearly, since $e_3(x_0) \neq 0$ for the starting procedure (9.1c), the sequence (9.3) diverges at x_0 like $\mathcal{O}(1/h)$ in this case.

We conclude from this example that for linear multistep methods there is in general no asymptotic expansion of the form

$$y_n - y(x_n) = e_p(x_n)h^p + e_{p+1}(x_n)h^{p+1} + \dots$$

which holds uniformly for $0 \leq nh \leq \text{Const}$. It will be necessary to add perturbation terms

$$y_n - y(x_n) = (e_p(x_n) + \varepsilon_n^p)h^p + (e_{p+1}(x_n) + \varepsilon_n^{p+1})h^{p+1} + \dots \quad (9.4)$$

which compensate the irregularity near x_0 . If the perturbations ε_n^j decay exponentially (for $n \rightarrow \infty$), then they have no influence on the asymptotic expansion for x_n bounded away from x_0 .

Asymptotic Expansion for Strictly Stable Methods (8.4)

In order to extend the techniques of Section II.8 to multistep methods it is useful to write them as a “one-step” method in a higher dimensional space (cf. (4.8) and Example 8.2). This suggests we study at once the asymptotic expansion for the general method (8.4). Because of the presence of $\varepsilon_n^j h^j$ in (9.4), the iterative proof of Theorem 9.1 below will lead us to increment functions which also depend on n , of the form

$$\Phi_n(x, u, h) = \Phi(x, u + h\alpha_n(h), h) + \beta_n(h). \quad (9.5)$$

We therefore consider for an equidistant grid (x_n) the numerical procedure

$$\begin{aligned} u_0 &= \varphi(h) \\ u_{n+1} &= Su_n + h\Phi_n(x_n, u_n, h), \end{aligned} \quad (9.6)$$

where Φ_n is given by (9.5) and the correct value function is again denoted by $z(x, h)$. The following additional assumptions will simplify the discussion of an asymptotic expansion:

- A1) Method (9.6) is *strictly stable*; i.e., it is stable (Definition 8.8) and 1 is the only eigenvalue of S with modulus one. In this case the spectral radius of $S - E$ (cf. formula (8.11)) is smaller than 1;
- A2) $\alpha_n(h)$ and $\beta_n(h)$ are polynomials, whose coefficients *decay exponentially* like $\mathcal{O}(\varrho_0^n)$ for $n \rightarrow \infty$. Here ϱ_0 denotes some number lying between the spectral radius of $S - E$ and one; i.e. $\varrho(S - E) < \varrho_0 < 1$;
- A3) the functions φ , z and Φ are sufficiently differentiable.

Assumption A3 allows us to expand the local error, defined by (8.9), into a Taylor series:

$$\begin{aligned} d_{n+1} &= z(x_n + h, h) - Sz(x_n, h) - h\Phi(x_n, z(x_n, h) + h\alpha_n(h), h) - h\beta_n(h) \\ &= d_0(x_n) + d_1(x_n)h + \dots + d_{N+1}(x_n)h^{N+1} \\ &\quad - h^2 \frac{\partial \Phi}{\partial u}(x_n, z(x_n, 0), 0)\alpha_n(h) - \dots - h\beta_n(h) + \mathcal{O}(h^{N+1}). \end{aligned}$$

The expressions involving $\alpha_n(h)$ can be simplified further. Indeed, for a smooth function $G(x)$ we have

$$G(x_n)\alpha_n(h) = G(x_0)\alpha_n(h) + hG'(x_0)n\alpha_n(h) + \dots + h^{N+1}R(n, h).$$

We observe that $n^j\alpha_n(h)$ is again a polynomial in h and that its coefficients decay like $\mathcal{O}(\varrho^n)$ where ϱ satisfies $\varrho_0 < \varrho < 1$. The same argument shows the boundedness of the remainder $R(n, h)$ for $0 \leq nh \leq \text{Const}$. As a consequence we can

write the local error in the form

$$\begin{aligned} d_0 &= \gamma_0 + \gamma_1 h + \dots + \gamma_N h^N + \mathcal{O}(h^{N+1}) \\ d_{n+1} &= (d_0(x_n) + \delta_n^0) + \dots + (d_{N+1}(x_n) + \delta_n^{N+1}) h^{N+1} + \mathcal{O}(h^{N+2}) \quad (9.7) \\ &\text{for } 0 \leq nh \leq \text{Const.} \end{aligned}$$

The functions $d_j(x)$ are smooth and the perturbations δ_n^j satisfy $\delta_n^j = \mathcal{O}(\varrho^n)$. The expansion (9.7) is unique, because $\delta_n^j \rightarrow 0$ for $n \rightarrow \infty$.

Method (9.6) is called *consistent of order p* , if the local error (9.7) satisfies (Lemma 8.11)

$$d_n = \mathcal{O}(h^p) \quad \text{for } 0 \leq nh \leq \text{Const}, \quad \text{and} \quad Ed_p(x) = 0. \quad (9.8)$$

Observe that by this definition the perturbations δ_n^j have to vanish for $j = 0, \dots, p-1$, but no condition is imposed on δ_n^p . The exponential decay of these terms implies that we still have

$$d_{n+1} + E(d_n + \dots + d_0) = \mathcal{O}(h^p) \quad \text{for } 0 \leq nh \leq \text{Const},$$

in agreement with Definition 8.10. One can now easily verify that Lemma 8.12 (Φ_n satisfies a Lipschitz condition with the same constant as Φ) and the Convergence Theorem 8.13 remain valid for method (9.6). In the following theorem we use, as for one-step methods, the notation $u_h(x) = u_n$ when $x = x_n$.

Theorem 9.1 (Hairer & Lubich 1984). *Let the method (9.6) satisfy A1-A3 and be consistent of order $p \geq 1$. Then the global error has an asymptotic expansion of the form*

$$u_h(x) - z(x, h) = e_p(x)h^p + \dots + e_N(x)h^N + E(x, h)h^{N+1} \quad (9.9)$$

where the $e_j(x)$ are given in the proof (cf. formula (9.18)) and $E(x, h)$ is bounded uniformly in $h \in [0, h_0]$ and for x in compact intervals not containing x_0 . More precisely than (9.9), there is an expansion

$$u_n - z_n = (e_p(x_n) + \varepsilon_n^p)h^p + \dots + (e_N(x_n) + \varepsilon_n^N)h^N + \tilde{E}(n, h)h^{N+1} \quad (9.10)$$

where $\varepsilon_n^j = \mathcal{O}(\varrho^n)$ with $\varrho(S - E) < \varrho < 1$ and $\tilde{E}(n, h)$ is bounded for $0 \leq nh \leq \text{Const}$.

Remark. We obtain from (9.10) and (9.9)

$$E(x_n, h) = \tilde{E}(n, h) + h^{-1}\varepsilon_n^N + h^{-2}\varepsilon_n^{N-1} + \dots + h^{p-N-1}\varepsilon_n^p,$$

so that the remainder term $E(x, h)$ is in general not uniformly bounded in h for x varying in an interval $[x_0, \bar{x}]$. However, if x is bounded away from x_0 , say $x \geq x_0 + \delta$ ($\delta > 0$ fixed), the sequence ε_n^j goes to zero faster than any power of $\delta/n \leq h$.

Proof. a) As for one-step methods (cf. proof of Theorem 8.1, Chapter II) we construct a new method, which has as numerical solution

$$\hat{u}_n = u_n - (e(x_n) + \varepsilon_n)h^p \quad (9.11)$$

for a given smooth function $e(x)$ and a given sequence ε_n satisfying $\varepsilon_n = \mathcal{O}(\varrho^n)$. Such a method is given by

$$\begin{aligned} \hat{u}_0 &= \hat{\varphi}(h) \\ \hat{u}_{n+1} &= S\hat{u}_n + h\hat{\Phi}_n(x_n, \hat{u}_n, h) \end{aligned} \quad (9.12)$$

where $\hat{\varphi}(h) = \varphi(h) - (e(x_0) + \varepsilon_0)h^p$ and

$$\begin{aligned} \hat{\Phi}_n(x, u, h) &= \Phi_n(x, u + (e(x) + \varepsilon_n)h^p, h) \\ &\quad - (e(x+h) - Se(x))h^{p-1} - (\varepsilon_{n+1} - S\varepsilon_n)h^{p-1}. \end{aligned}$$

Since Φ_n is of the form (9.5), $\hat{\Phi}_n$ is also of this form, so that its local error has an expansion (9.7). We shall now determine $e(x)$ and ε_n in such a way that the method (9.12) is consistent of order $p+1$.

b) The local error \hat{d}_n of (9.12) can be expanded as

$$\begin{aligned} \hat{d}_0 &= z_0 - \hat{u}_0 = (\gamma_p + e(x_0) + \varepsilon_0)h^p + \mathcal{O}(h^{p+1}) \\ \hat{d}_{n+1} &= z_{n+1} - S\hat{u}_n - h\hat{\Phi}_n(x_n, \hat{u}_n, h) \\ &= d_{n+1} + ((I-S)e(x_n) + (\varepsilon_{n+1} - S\varepsilon_n))h^p \\ &\quad + (-G(x_n)(e(x_n) + \varepsilon_n) + e'(x_n))h^{p+1} + \mathcal{O}(h^{p+2}). \end{aligned}$$

Here

$$G(x) = \frac{\partial \Phi_n}{\partial u}(x, z(x, 0), 0)$$

which is independent of n by (9.5). The method (9.12) is consistent of order $p+1$, if (see (9.8))

- i) $\varepsilon_0 = -\gamma_p - e(x_0)$,
- ii) $d_p(x) + (I-S)e(x) + \delta_n^p + \varepsilon_{n+1} - S\varepsilon_n = 0 \quad \text{for } x = x_n$,
- iii) $Ee'(x) = EG(x)e(x) - Ed_{p+1}(x)$.

We assume for the moment that the system (i)-(iii) can be solved for $e(x)$ and ε_n . This will actually be demonstrated in part (d) of the proof. By the Convergence Theorem 8.13 the method (9.12) is convergent of order $p+1$. Hence

$$\hat{u}_n - z_n = \mathcal{O}(h^{p+1}) \quad \text{uniformly for } 0 \leq nh \leq \text{Const},$$

which yields the statement (9.10) for $N = p$.

c) The method (9.12) satisfies the assumptions of the theorem with p replaced by $p+1$ and ϱ_0 by ϱ . As in Theorem 8.1 (Section II.8) an induction argument yields the result.

d) It remains to find a solution of the system (i)-(iii). Condition (ii) is satisfied if

$$(iia) \quad d_p(x) = (S - I)(e(x) + c)$$

$$(iib) \quad \varepsilon_{n+1} - c = S(\varepsilon_n - c) - \delta_n^p$$

hold for some constant c . Using $(I - S + E)^{-1}(I - S) = (I - E)$, which is a consequence of $SE = E^2 = E$ (see (8.11)), formula (iia) is equivalent to

$$(I - S + E)^{-1}d_p(x) = -(I - E)(e(x) + c). \quad (9.13)$$

From (i) we obtain $\varepsilon_0 - c = -\gamma_p - (e(x_0) + c)$, so that by (9.13)

$$(I - E)(\varepsilon_0 - c) = -(I - E)\gamma_p + (I - S + E)^{-1}d_p(x_0).$$

Since $Ed_p(x_0) = 0$, this relation is satisfied in particular if

$$\varepsilon_0 - c = -(I - E)\gamma_p + (I - S + E)^{-1}d_p(x_0). \quad (9.14)$$

The numbers $\varepsilon_n - c$ are now determined by the recurrence relation (iib)

$$\begin{aligned} \varepsilon_n - c &= S^n(\varepsilon_0 - c) - \sum_{j=1}^n S^{n-j}\delta_{j-1}^p \\ &= E(\varepsilon_0 - c) + (S - E)^n(\varepsilon_0 - c) - E \sum_{j=0}^{\infty} \delta_j^p + E \sum_{j=n}^{\infty} \delta_j^p - \sum_{j=1}^n (S - E)^{n-j}\delta_{j-1}^p, \end{aligned}$$

where we have used $S^n = E + (S - E)^n$. If we put

$$c = E \sum_{j=0}^{\infty} \delta_j^p \quad (9.15)$$

the sequence $\{\varepsilon_n\}$ defined above satisfies $\varepsilon_n = \mathcal{O}(\varrho^n)$, since $E(\varepsilon_0 - c) = 0$ by (9.14) and since $\delta_n^p = \mathcal{O}(\varrho^n)$.

In order to find $e(x)$ we define

$$v(x) = Ee(x).$$

With the help of formulas (9.15) and (9.13) we can recover $e(x)$ from $v(x)$ by

$$e(x) = v(x) - (I - S + E)^{-1}d_p(x). \quad (9.16)$$

Equation (iii) can now be rewritten as the differential equation

$$v'(x) = EG(x) \left(v(x) - (I - S + E)^{-1}d_p(x) \right) - Ed_{p+1}(x), \quad (9.17)$$

and condition (i) yields the starting value $v(x_0) = -E(\gamma_p + \varepsilon_0)$. This initial value problem can be solved for $v(x)$ and we obtain $e(x)$ by (9.16). This function and the ε_n defined above represent a solution of (i)-(iii). \square

Remarks. a) It follows from (9.15)-(9.17) that the principal error term satisfies

$$\begin{aligned} e'_p(x) &= EG(x)e_p(x) - Ed_{p+1}(x) - (I - S + E)^{-1}d'_p(x) \\ e_p(x_0) &= -E\gamma_p - E \sum_{j=0}^{\infty} \delta_j^p - (I - S + E)^{-1}d_p(x_0). \end{aligned} \quad (9.18)$$

b) Since $e_{p+1}(x)$ is just the principal error term of method (9.12), it satisfies the differential equation (9.18) with d_j replaced by \hat{d}_{j+1} . By an induction argument we therefore have for $j \geq p$

$$e'_j(x) = EG(x)e_j(x) + \text{inhomogeneity}(x).$$

Weakly Stable Methods

We next study the asymptotic expansion for stable methods, which are not strictly stable. For example, the explicit mid-point rule (1.13'), treated in connection with the GBS-algorithm (Section II.9), is of this type. As at the beginning of this section, we apply the mid-point rule to the problem $y' = -y$, $y(0) = 1$ and consider the following three starting procedures

$$y_0 = 1, \quad y_1 = \exp(-h) \quad (9.19a)$$

$$y_0 = 1, \quad y_1 = 1 - h + \frac{h^2}{2} \quad (9.19b)$$

$$y_0 = 1, \quad y_1 = 1 - h. \quad (9.19c)$$

The three pictures on the left of Fig. 9.2 show the global error divided by h^2 . For the first two starting procedures we have convergence to the function $xe^{-x}/6$, while for (9.19c) the divided error $(y_n - y(x_n))/h^2$ converges to

$$\begin{aligned} e^{-x} \left(\frac{2x-3}{12} \right) + \frac{e^x}{4} & \quad \text{for } n \text{ even,} \\ e^{-x} \left(\frac{2x-3}{12} \right) - \frac{e^x}{4} & \quad \text{for } n \text{ odd.} \end{aligned}$$

We then subtract the h^2 -term from the global error and divide by h^3 in the case (9.19a) and by h^4 for (b) and (c). The result is plotted in the pictures on the right of Fig. 9.2.

This example nicely illustrates the fact that we no longer have an asymptotic expansion of the form (9.9) or (9.10) but that there exists one expansion for x_n with n even, and a different expansion for x_n with n odd (see also Exercise 2 of Section II.9). Similar results for more general methods will be obtained here.

We say that a method of the form (8.4) is *weakly stable*, if it is stable, but if the matrix S has, besides $\zeta_1 = 1$, further eigenvalues of modulus 1, say ζ_2, \dots, ζ_l .

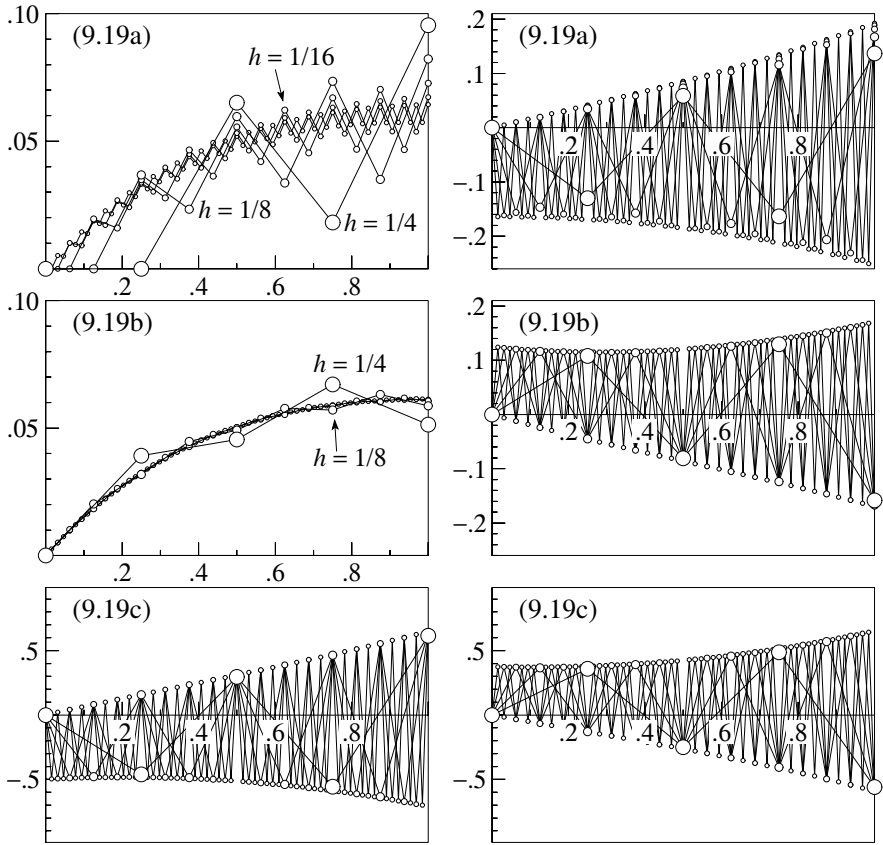


Fig. 9.2. Asymptotic expansion of the mid-point rule
(three different starting procedures)

The matrix S therefore has the representation (cf. (8.11))

$$S = \zeta_1 E_1 + \zeta_2 E_2 + \dots + \zeta_l E_l + R \quad (9.20)$$

where the E_j are the projectors (corresponding to ζ_j) and the spectral radius of R satisfies $\varrho(R) < 1$.

In what follows we restrict ourselves to the case where all ζ_j ($j = 1, \dots, l$) are roots of unity. This allows a simple proof for the existence of an asymptotic expansion and is at the same time by far the most important special case. For the general situation we refer to Hairer & Lubich (1984).

Theorem 9.2. *Let the method (9.6) with Φ_n independent of n be stable, consistent of order p and satisfy A3. If all eigenvalues (of S) of modulus 1 satisfy $\zeta_j^q = 1$ ($j = 1, \dots, l$) for some positive integer q , then we have an asymptotic expansion*

of the form ($\omega = e^{2\pi i/q}$)

$$u_n - z_n = \sum_{s=0}^{q-1} \omega^{ns} \left(e_{ps}(x_n) h^p + \dots + e_{Ns}(x_n) h^N \right) + E(n, h) h^{N+1} \quad (9.21)$$

where the $e_{js}(x)$ are smooth functions and $E(n, h)$ is uniformly bounded for $0 < \delta \leq nh \leq \text{Const}$.

Proof. The essential idea of the proof is to consider q consecutive steps of method (9.6) as one method over a large step. Putting $\tilde{u}_n = u_{nq+i}$ ($0 \leq i \leq q-1$ fixed), $\tilde{h} = qh$ and $\tilde{x}_n = x_i + n\tilde{h}$, this method becomes

$$\tilde{u}_{n+1} = S^q \tilde{u}_n + \tilde{h} \tilde{\Phi}(\tilde{x}_n, \tilde{u}_n, \tilde{h}) \quad (9.22)$$

with a suitably chosen $\tilde{\Phi}$. E.g., for $q = 2$ we have

$$\tilde{\Phi}(\tilde{x}, \tilde{u}, \tilde{h}) = \frac{1}{2} S \Phi\left(\tilde{x}, \tilde{u}, \frac{\tilde{h}}{2}\right) + \frac{1}{2} \Phi\left(\tilde{x} + \frac{\tilde{h}}{2}, S\tilde{u} + \frac{\tilde{h}}{2} \Phi\left(\tilde{x}, \tilde{u}, \frac{\tilde{h}}{2}\right), \frac{\tilde{h}}{2}\right).$$

The assumption on the eigenvalues implies

$$S^q = E_1 + \dots + E_l + R^q$$

so that (9.22) is seen to be a strictly stable method. A straightforward calculation shows that the local error of (9.22) satisfies

$$\begin{aligned} \tilde{d}_0 &= \mathcal{O}(h^p) \\ \tilde{d}_{n+1} &= (I + S + \dots + S^{q-1}) d_p(\tilde{x}_n) h^p + \mathcal{O}(h^{p+1}). \end{aligned}$$

Inserting (9.20) and using $\zeta_j^q = 1$ we obtain, with $\tilde{E} = E_1 + \dots + E_l$,

$$\begin{aligned} &\tilde{E}(I + S + \dots + S^{q-1}) d_p(x) \\ &= \tilde{E} \left(I - \tilde{E} + qE_1 + \sum_{j=2}^l \frac{1 - \zeta_j^q}{1 - \zeta_j} E_j + \sum_{j=1}^{q-1} R^j \right) d_p(x) = qE_1 d_p(x), \end{aligned}$$

which vanishes by (8.15). Hence, also method (9.22) is consistent of order p . All the assumptions of Theorem 9.1 are thus verified for method (9.22). We therefore obtain

$$u_{nq+i} - z_{nq+i} = \tilde{e}_{pi}(x_{nq+i}) h^p + \dots + \tilde{e}_{Ni}(x_{nq+i}) h^N + E_i(n, h) h^{N+1}$$

where $E_i(n, h)$ has the desired boundedness properties. If we define $e_{js}(x)$ as a solution of the Vandermonde-type system

$$\sum_{s=0}^{q-1} \omega^{is} e_{js}(x) = \tilde{e}_{ji}(x)$$

we obtain (9.21). □

The Adjoint Method

For a method (8.4) the correct value function $z(x, h)$, the starting procedure $\varphi(h)$ and the increment function $\Phi(x, u, h)$ are usually also defined for negative h (see the examples of Section III.8). As for one-step methods (Section II.8) we shall give here a precise meaning to the numerical solution $u_h(x)$ for negative h . This then leads in a natural way to the study of asymptotic expansions in even powers of h .

With the notation $u_h(x) = u_n$ for $x = x_0 + nh$ ($h > 0$) the method (8.4) becomes

$$\begin{aligned} u_h(x_0) &= \varphi(h) \\ u_h(x+h) &= Su_h(x) + h\Phi(x, u_h(x), h) \quad \text{for } x = x_0 + nh. \end{aligned} \quad (9.23)$$

We first replace h by $-h$ in (9.23) to obtain

$$\begin{aligned} u_{-h}(x_0) &= \varphi(-h) \\ u_{-h}(x-h) &= Su_{-h}(x) - h\Phi(x, u_{-h}(x), -h) \end{aligned}$$

and then x by $x+h$ which gives

$$\begin{aligned} u_{-h}(x_0) &= \varphi(-h) \\ u_{-h}(x) &= Su_{-h}(x+h) - h\Phi(x+h, u_{-h}(x+h), -h). \end{aligned}$$

For sufficiently small h this equation can be solved for $u_{-h}(x+h)$ (Implicit Function Theorem) and we obtain

$$\begin{aligned} u_{-h}(x_0) &= \varphi(-h), \\ u_{-h}(x+h) &= S^{-1}u_{-h}(x) + h\Phi^*(x, u_{-h}(x), h). \end{aligned} \quad (9.24)$$

The method (9.24), which is again of the form (8.4), is called the *adjoint method* of (9.23). Its correct value function is $z^*(x, h) = z(x, -h)$. Observe that for given S and Φ the new increment function Φ^* is just defined by the pair of formulas

$$\begin{aligned} v &= Su - h\Phi(x+h, u, -h) \\ u &= S^{-1}v + h\Phi^*(x, v, h). \end{aligned} \quad (9.25)$$

Example 9.3. Consider a linear multistep method with generating functions

$$\varrho(\zeta) = \sum_{j=0}^k \alpha_j \zeta^j, \quad \sigma(\zeta) = \sum_{j=0}^k \beta_j \zeta^j.$$

Then we have

$$S = \begin{pmatrix} -\alpha_{k-1}/\alpha_k & -\alpha_{k-2}/\alpha_k & \cdots & -\alpha_0/\alpha_k \\ 1 & 0 & \cdots & 0 \\ & 1 & \ddots & 0 \\ & & \vdots & \vdots \\ & & & 1 & 0 \end{pmatrix}, \quad \Phi(x, u, h) = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \psi(x, u, h)$$

where $\psi = \psi(x, u, h)$ is the solution of $(u = (u_{k-1}, \dots, u_0)^T)$

$$\alpha_k \psi = \sum_{j=0}^{k-1} \beta_j f(x + jh, u_j) + \beta_k f\left(x + kh, h\psi - \sum_{j=0}^{k-1} \frac{\alpha_j}{\alpha_k} u_j\right).$$

A straightforward use of the formulas (9.25) shows that

$$S^{-1} = \begin{pmatrix} 0 & 1 & & \\ 0 & 0 & & \\ \vdots & \vdots & \dots & 1 \\ -\alpha_k/\alpha_0 & -\alpha_{k-1}/\alpha_0 & \dots & -\alpha_1/\alpha_0 \end{pmatrix}, \quad \Phi^*(x, v, h) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \psi^*(x, v, h)$$

where $\psi^* = \psi^*(x, v, h)$ (with $v = (v_0, \dots, v_{k-1})^T$) is given by

$$-\alpha_0 \psi^* = \sum_{j=0}^{k-1} \beta_{k-j} f(x + (j - k + 1)h, v_j) + \beta_0 f\left(x + h, h\psi^* - \sum_{j=0}^{k-1} \frac{\alpha_{k-j}}{\alpha_0} v_j\right).$$

This shows that the adjoint method is again a linear multistep method. Its generating polynomials are

$$\varrho^*(\zeta) = -\zeta^k \varrho(\zeta^{-1}), \quad \sigma^*(\zeta) = \zeta^k \sigma(\zeta^{-1}). \quad (9.26)$$

Our next aim is to prove that the adjoint method has exactly the same asymptotic expansion as the original method, with h replaced by $-h$. For this it is necessary that S^{-1} also be a stable matrix. Therefore all eigenvalues of S must lie on the unit circle.

Theorem 9.4. *Let the method (9.23) be stable, consistent of order p and assume that all eigenvalues of S satisfy $\zeta_j^q = 1$ for some positive integer q . Then the global error has an asymptotic expansion of the form $(\omega = e^{2\pi i/q})$*

$$u_h(x_n) - z(x_n, h) = \sum_{s=0}^{q-1} \omega^{ns} \left(e_{ps}(x_n) h^p + \dots + e_{Ns}(x_n) h^N \right) + E(x_n, h) h^{N+1}, \quad (9.27)$$

valid for positive and negative h . The remainder $E(x, h)$ is uniformly bounded for $|h| \leq h_0$ and $x_0 \leq x \leq \hat{x}$.

Proof. As in the proof of Theorem 9.2 we consider q consecutive steps of method (9.23) as one new method. The assumption on the eigenvalues implies that $S^q = I = \text{identity}$. Therefore the new method is essentially a one-step method. The only difference is that here the starting procedure and the correct value function may depend on h . A straightforward extension of Theorem 8.5 of Chapter II (Exercise 3) implies the existence of an expansion

$$\begin{aligned} u_h(x_{nq+i}) - z(x_{nq+i}, h) &= \tilde{e}_{pi}(x_{nq+i}) h^p + \dots + \tilde{e}_{Ni}(x_{nq+i}) h^N \\ &\quad + E_i(x_{nq+i}, h) h^{N+1}. \end{aligned}$$

This expansion is valid for positive and negative h ; the remainder $E_i(x, h)$ is bounded for $|h| \leq h_0$ and $x_0 \leq x \leq \hat{x}$. The same argument as in the proof of Theorem 9.2 now leads to the desired expansion. \square

Symmetric Methods

The definition of symmetry for general linear methods is not as straightforward as for one-step methods. In Example 9.3 we saw that the components of the numerical solution of the adjoint method are in inverse order. Therefore, it is too restrictive to require that $\varphi(h) = \varphi(-h)$, $S = S^{-1}$ and $\Phi = \Phi^*$.

However, for many methods of practical interest the correct value function satisfies a *symmetry relation* of the form

$$z(x, h) = Qz(x + qh, -h) \quad (9.28)$$

where Q is a square matrix and q an integer. This is for instance the case for linear multistep methods, where the correct value function is given by

$$z(x, h) = (y(x + (k-1)h), \dots, y(x))^T.$$

The relation (9.28) holds with

$$Q = \begin{pmatrix} & & 1 \\ & \ddots & \\ 1 & & \end{pmatrix} \quad \text{and} \quad q = k-1. \quad (9.29)$$

Definition 9.5. Suppose that the correct value function satisfies (9.28). Method (9.23) is called *symmetric* (with respect to (9.28)), if the numerical solution satisfies its analogue

$$u_h(x) = Qu_{-h}(x + qh). \quad (9.30)$$

Example 9.6. Consider a linear multistep method and suppose that the generating polynomials of the adjoint method (9.26) satisfy

$$\varrho^*(\zeta) = \varrho(\zeta), \quad \sigma^*(\zeta) = \sigma(\zeta). \quad (9.31)$$

This is equivalent to the requirement (cf. (3.24))

$$\alpha_{k-j} = -\alpha_j, \quad \beta_{k-j} = \beta_j.$$

A straightforward calculation (using the formulas of Example 9.3) then shows that the symmetry relation (9.30) holds for all $x = x_0 + nh$ whenever it holds for $x = x_0$. This imposes an additional condition on the starting procedure $\varphi(h)$.

Let us finally demonstrate how Theorem 9.4 can be used to prove *asymptotic expansions in even powers of h* . Denote by $u_h^j(x)$ the j th component of $u_h(x)$. The symmetry relation (9.30) for multistep methods then implies

$$u_{-h}^k(x) = u_h^1(x - (k-1)h)$$

Furthermore, for any multistep method we have

$$u_h^k(x) = u_h^1(x - (k-1)h)$$

so that

$$u_h^k(x) = u_{-h}^k(x)$$

for symmetric methods. As a consequence of Theorem 9.4 the asymptotic expansion of the global error is in even powers of h , whenever the multistep method is symmetric in the sense of Definition 9.5.

Exercises

1. Consider a strictly stable, p th order, linear multistep method written in the form (9.6) (see Example 9.3) and set

$$G(x) = \frac{\partial \Phi}{\partial u}(x, z(x, 0), 0).$$

- a) Prove that

$$EG(x)\mathbb{1} = \mathbb{1} \frac{\partial f}{\partial y}(x, y(x))$$

where E is the matrix given by (8.11) and $\mathbb{1} = (1, \dots, 1)^T$.

- b) Show that the function $e_p(x)$ in the expansion (9.9) is given by $e_p(x) = \mathbb{1}\hat{e}_p(x)$, where

$$\hat{e}_p(x) = \frac{\partial f}{\partial y}(x, y(x))\hat{e}_p(x) - Cy^{(p+1)}(x)$$

and C is the error constant (cf. (2.13)). Compute also $\hat{e}_p(x_0)$.

2. For the 3-step BDF-method, applied to $y' = -y$, $y(0) = 1$ with starting procedure (9.1c), compute the function $e_3(x)$ and the perturbations $\{\varepsilon_n^3\}_{n \geq 0}$ in the expansion (9.4). Compare your result with Fig. 9.1.

3. Consider the method

$$u_0 = \varphi(h), \quad u_{n+1} = u_n + h\Phi(x_n, u_n, h) \quad (9.32)$$

with correct value function $z(x, h)$.

- a) Prove that the global error has an asymptotic expansion of the form

$$u_n - z_n = e_p(x_n)h^p + \dots + e_N(x_n)h^N + E(x_n, h)h^{N+1}$$

where $E(x, h)$ is uniformly bounded for $0 \leq h \leq h_0$ and $x_0 \leq x \leq \hat{x}$.

- b) Show that Theorem 8.5 of Chapter II remains valid for method (9.32).

III.10 Multistep Methods for Second Order Differential Equations

En 1904 j'eus besoin d'une pareille méthode pour calculer les trajectoires des corpuscules électrisés dans un champ magnétique, et en essayant diverses méthodes déjà connues, mais sans les trouver assez commodes pour mon but, je fus conduit moi-même à élaborer une méthode assez simple, dont je me suis servi ensuite.

(C. Störmer 1921)

Because of their importance, second order differential equations deserve some additional attention. We already saw in Section II.14 that for special second order differential equations certain direct one-step methods are more efficient than the classical Runge-Kutta methods. We now investigate whether a similar situation also holds for multistep methods.

Consider the second order differential equation

$$y'' = f(x, y, y') \quad (10.1)$$

where y is allowed to be a vector. We rewrite (10.1) in the usual way as a first order system and apply a multistep method

$$\begin{aligned} \sum_{i=0}^k \alpha_i y_{n+i} &= h \sum_{i=0}^k \beta_i y'_{n+i} \\ \sum_{i=0}^k \alpha_i y'_{n+i} &= h \sum_{i=0}^k \beta_i f(x_{n+i}, y_{n+i}, y'_{n+i}). \end{aligned} \quad (10.2)$$

If the right hand side of the differential equation does not depend on y' ,

$$y'' = f(x, y), \quad (10.3)$$

it is natural to look for numerical methods which do not involve the first derivative. An elimination of $\{y'_n\}$ in the equations (10.2) results in

$$\sum_{i=0}^{2k} \hat{\alpha}_i y_{n+i} = h^2 \sum_{i=0}^{2k} \hat{\beta}_i f(x_{n+i}, y_{n+i}) \quad (10.4)$$

where the new coefficients $\hat{\alpha}_i, \hat{\beta}_i$ are given by

$$\sum_{i=0}^{2k} \hat{\alpha}_i \zeta^i = \left(\sum_{i=0}^k \alpha_i \zeta^i \right)^2, \quad \sum_{i=0}^{2k} \hat{\beta}_i \zeta^i = \left(\sum_{i=0}^k \beta_i \zeta^i \right)^2. \quad (10.5)$$

In what follows we investigate (10.4) with coefficients that do not necessarily satisfy (10.5). It is hoped to achieve the same order with a smaller step number.

Explicit Störmer Methods

Sein Vortrag ist übrigens ziemlich trocken und langweilig . . .
(B. Riemann's opinion about Encke, 1847)

Had the Ast. Ges. Essay been entirely free from numerical blunders, . . .
(P.H. Cowell & A.C.D. Crommelin 1910)

Since most differential equations of celestial mechanics are of the form (10.3) it is not surprising that the first attempts at developing special methods for these equations were made by astronomers.

For his extensive numerical calculations concerning the aurora borealis (see below), C. Störmer (1907) developed an accurate and simple method as follows: by adding the Taylor series for $y(x_n + h)$ and $y(x_n - h)$ we obtain

$$y(x_n + h) - 2y(x_n) + y(x_n - h) = h^2 y''(x_n) + \frac{h^4}{12} y^{(4)}(x_n) + \frac{h^6}{360} y^{(6)}(x_n) + \dots$$

If we insert $y''(x_n)$ from the differential equation (10.3) and neglect higher terms, we get

$$y_{n+1} - 2y_n + y_{n-1} = h^2 f_n$$

as a first simple method, which is sometimes called Störmer's or Encke's method. For greater precision, we replace the higher derivatives of y by central differences of f

$$\begin{aligned} h^2 y^{(4)}(x_n) &= \Delta^2 f_{n-1} - \frac{1}{12} \Delta^4 f_{n-2} + \dots \\ h^4 y^{(6)}(x_n) &= \Delta^4 f_{n-2} + \dots \end{aligned}$$

and obtain

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \left(f_n + \frac{1}{12} \Delta^2 f_{n-1} - \frac{1}{240} \Delta^4 f_{n-2} + \dots \right). \quad (10.6)$$

This formula is not yet very practical, since the differences of the right hand side contain the unknown expressions f_{n+1} and f_{n+2} . Neglecting fifth-order differences (i.e., putting $\Delta^4 f_{n-2} \approx \Delta^4 f_{n-4}$ and $\Delta^2 f_{n-1} = \Delta^2 f_{n-2} + \Delta^3 f_{n-3} + \Delta^4 f_{n-3} \approx \Delta^2 f_{n-2} + \Delta^3 f_{n-3} + \Delta^4 f_{n-4}$) one gets

$$y_{n+1} - 2y_n + y_{n-1} = h^2 f_n + \frac{h^2}{12} \left(\Delta^2 f_{n-2} + \Delta^3 f_{n-3} + \frac{19}{20} \Delta^4 f_{n-4} \right) \quad (10.7)$$

(“... formule qui est fondamentale dans notre méthode . . .”, C. Störmer 1907).

Some years later Cowell & Crommelin (1910) used the same ideas to investigate the motion of Halley's comet. They considered one additional term in the series (10.6), namely

$$\frac{31}{60480} \Delta^6 f_{n-3} \approx \frac{1}{1951} \Delta^6 f_{n-3}.$$

Arbitrary orders. Integrating equation (10.3) twice we obtain

$$y(x+h) = y(x) + hy'(x) + h^2 \int_0^1 (1-s)f(x+sh, y(x+sh)) ds. \quad (10.8)$$

In order to eliminate the first derivative of $y(x)$ we write the same formula with h replaced by $-h$ and add the two expressions:

$$\begin{aligned} y(x+h) - 2y(x) + y(x-h) \\ = h^2 \int_0^1 (1-s) \left(f(x+sh, y(x+sh)) + f(x-sh, y(x-sh)) \right) ds. \end{aligned} \quad (10.9)$$

As in the derivation of the Adams formulas (Section III.1) we replace the unknown function $f(t, y(t))$ by the interpolation polynomial $p(t)$ of formula (1.4). This yields the *explicit* method

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{j=0}^{k-1} \sigma_j \nabla^j f_n \quad (10.10)$$

with coefficients σ_j given by

$$\sigma_j = (-1)^j \int_0^1 (1-s) \left(\binom{-s}{j} + \binom{s}{j} \right) ds. \quad (10.11)$$

See Table 10.1 for their numerical values and Exercise 2 for their computation.

Table 10.1. Coefficients of the method (10.10)

j	0	1	2	3	4	5	6	7	8	9
σ_j	1	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{19}{240}$	$\frac{3}{40}$	$\frac{863}{12096}$	$\frac{275}{4032}$	$\frac{33953}{518400}$	$\frac{8183}{129600}$

Special cases of (10.10) are

$$k=2: \quad y_{n+1} - 2y_n + y_{n-1} = h^2 f_n$$

$$k=3: \quad y_{n+1} - 2y_n + y_{n-1} = h^2 \left(\frac{13}{12} f_n - \frac{1}{6} f_{n-1} + \frac{1}{12} f_{n-2} \right) \quad (10.10')$$

$$k=4: \quad y_{n+1} - 2y_n + y_{n-1} = h^2 \left(\frac{7}{6} f_n - \frac{5}{12} f_{n-1} + \frac{1}{3} f_{n-2} - \frac{1}{12} f_{n-3} \right).$$

Method (10.10) with $k=5$ is formula (10.7), the method used by Störmer (1907, 1921), and for $k=6$ one obtains the method used by Cowell & Crommelin (1910). The simplest of these methods ($k=1$ or $k=2$) has been successfully applied as the basis of an extrapolation method (Section II.14, formula (14.32)).

Implicit Störmer Methods

The first terms of (10.6)

$$\begin{aligned} y_{n+1} - 2y_n + y_{n-1} &= h^2 \left(f_n + \frac{1}{12} \Delta^2 f_{n-1} \right) \\ &= \frac{h^2}{12} (f_{n+1} + 10f_n + f_{n-1}) \end{aligned} \quad (10.12)$$

form an implicit equation for y_{n+1} . This can either be used in a predictor-corrector fashion, or, as advocated by B. Numerov (1924, 1927), by solving this implicit nonlinear equation directly for y_{n+1} .

To obtain more accurate formulas, analogous to the implicit Adams methods, we use the interpolation polynomial $p^*(t)$ of (1.8), which passes through the additional point (x_{n+1}, f_{n+1}) . This yields the implicit method

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{j=0}^k \sigma_j^* \nabla^j f_{n+1}, \quad (10.13)$$

where the coefficients σ_j^* are defined by

$$\sigma_j^* = (-1)^j \int_0^1 (1-s) \left(\binom{-s+1}{j} + \binom{s+1}{j} \right) ds \quad (10.14)$$

and are given in Table 10.2 for $j \leq 9$.

Table 10.2. Coefficients of the implicit method (10.13)

j	0	1	2	3	4	5	6	7	8	9
σ_j^*	1	-1	$\frac{1}{12}$	0	$\frac{-1}{240}$	$\frac{-1}{240}$	$\frac{-221}{60480}$	$\frac{-19}{6048}$	$\frac{-9829}{3628800}$	$\frac{-407}{172800}$

Further methods can be derived by using the ideas of Nyström and Milne for first order equations. With the substitutions $h \rightarrow 2h$, $2s \rightarrow s$ and $x \rightarrow x - h$ formula (10.9) becomes

$$\begin{aligned} y(x+h) - 2y(x-h) + y(x-3h) &= h^2 \int_0^2 (2-s) \\ &\cdot \left(f(x+(s-1)h, y(x+(s-1)h)) + f(x-(s+1)h, y(x-(s+1)h)) \right) ds. \end{aligned} \quad (10.15)$$

If one replaces $f(t, y(t))$ by the polynomial $p(t)$ (respectively $p^*(t)$) one obtains the new classes of explicit (respectively implicit) methods.

Numerical Example

Nous avons calculé plus de 120 trajectoires différentes, travail immense qui a exigé plus de 4500 heures . . . Quand on est suffisamment exercé, on calcule environ trois points (R, z) par heure.
(C. Störmer 1907)

We choose the historical problem treated by Störmer in 1907: Störmer's aim was to confirm numerically the conjecture of Birkeland, who explained in 1896 the aurora borealis as being produced by electrical particles emanating from the sun and dancing in the earth's magnetic field. Suppose that an elementary magnet is situated at the origin with its axis along to the z -axis. The trajectory $(x(s), y(s), z(s))$ of an electrical particle in this magnetic field then satisfies

$$\begin{aligned}x'' &= \frac{1}{r^5}(3yzz' - (3z^2 - r^2)y') \\y'' &= \frac{1}{r^5}((3z^2 - r^2)x' - 3xzz') \\z'' &= \frac{1}{r^5}(3xzy' - 3y zx')\end{aligned}\tag{10.16}$$

where $r^2 = x^2 + y^2 + z^2$. Introducing the polar coordinates

$$x = R \cos \varphi, \quad y = R \sin \varphi \tag{10.17}$$

the system (10.16) becomes equivalent to

$$R'' = \left(\frac{2\gamma}{R} + \frac{R}{r^3}\right)\left(\frac{2\gamma}{R^2} + \frac{3R^2}{r^5} - \frac{1}{r^3}\right) \tag{10.18a}$$

$$z'' = \left(\frac{2\gamma}{R} + \frac{R}{r^3}\right)\frac{3Rz}{r^5} \tag{10.18b}$$

$$\varphi' = \left(\frac{2\gamma}{R} + \frac{R}{r^3}\right)\frac{1}{R} \tag{10.18c}$$

where now $r^2 = R^2 + z^2$ and γ is some constant arising from the integration of φ'' . The two equations (10.18a,b) constitute a second order differential equation of type (10.3), which can be solved numerically by the methods of this section. φ is then obtained by simple integration of (10.18c). Störmer found after long calculations that the initial values

$$\begin{aligned}R_0 &= 0.257453, & z_0 &= 0.314687, & \gamma &= -0.5, \\R'_0 &= \sqrt{Q_0} \cos u, & z'_0 &= \sqrt{Q_0} \sin u, & u &= 5\pi/4 \\r_0 &= \sqrt{R_0^2 + z_0^2}, & Q_0 &= 1 - (2\gamma/R_0 + R_0/r_0^3)^2\end{aligned}\tag{10.18d}$$

produce a specially interesting solution curve approaching very closely the North Pole. Fig. 10.1 shows 125 solution curves (in the x, y, z -space) with these and neighbouring initial values to give an impression of how an aurora borealis comes into being.

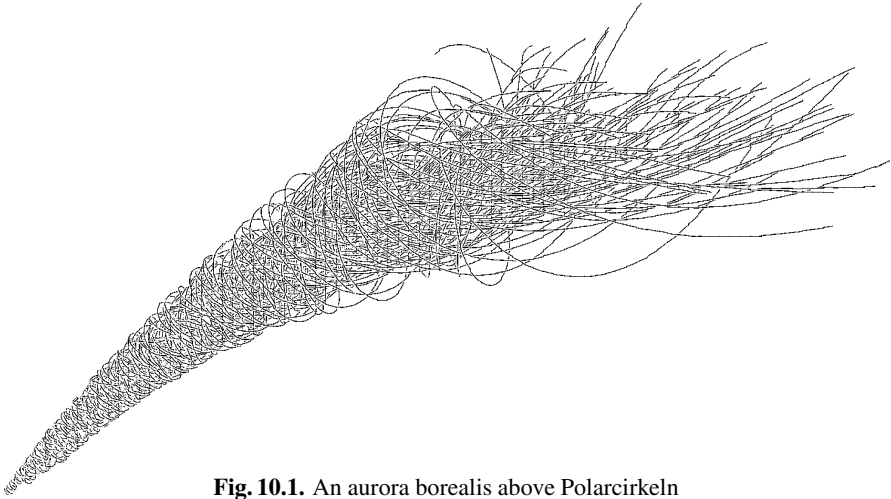


Fig. 10.1. An aurora borealis above Polarcirkeln

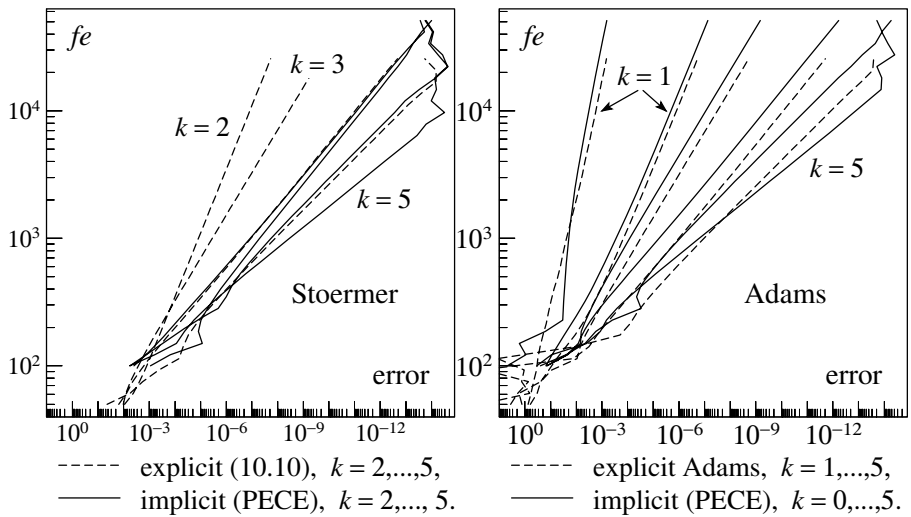


Fig. 10.2. Performance of Störmer and Adams methods

Fig. 10.2 compares the performance of the Störmer methods (10.10) and (10.13) (in PECE mode) with that of the Adams methods by integrating subsystem (10.18a,b) with initial values (10.18d) for $0 \leq s \leq 0.3$. The diagrams compare the Euclidean norm in \mathbb{R}^2 of the error of the final solution point (R, z) with the number of function evaluations fe . The step numbers used are $\{n = 50 \cdot 2^{0.3 \cdot i}\}_{i=0,1,\dots,30} = \{50, 61, 75, 93, 114, \dots, 25600\}$. The starting values were computed very precisely with an explicit Runge-Kutta method and step size $h_{RK} = h/10$. It can be observed that the Störmer methods are substantially more precise due to the smaller error constants (compare Tables 10.1 and 10.2 with Tables 1.1

and 1.2). In addition, they have lower overhead. However, they must be implemented carefully in order to avoid rounding errors (see below).

General Formulation

Our next aim is to study stability, consistency and convergence of general linear multistep methods for (10.3). We write them in the form

$$\sum_{i=0}^k \alpha_i y_{n+i} = h^2 \sum_{i=0}^k \beta_i f(x_{n+i}, y_{n+i}). \quad (10.19)$$

The generating polynomials of the coefficients α_i and β_i are again denoted by

$$\varrho(\zeta) = \sum_{i=0}^k \alpha_i \zeta^i, \quad \sigma(\zeta) = \sum_{i=0}^k \beta_i \zeta^i. \quad (10.20)$$

If we apply method (10.19) to the initial value problem

$$y'' = f(x, y), \quad y(x_0) = y_0, \quad y'(x_0) = y'_0 \quad (10.21)$$

it is natural to require that the starting values be consistent with both initial values, i.e., that

$$\frac{y_i - y_0 - i h y'_0}{h} \rightarrow 0 \quad \text{for } h \rightarrow 0, \quad i = 0, 1, \dots, k-1. \quad (10.22)$$

For the *stability condition* of method (10.19) we consider the simple problem

$$y'' = 0, \quad y_0 = 0, \quad y'_0 = 0.$$

Its numerical solution satisfies a linear difference equation with $\varrho(\zeta)$ as characteristic polynomial. The same considerations as in the proof of Theorem 4.2 show that the following stability condition is necessary for convergence.

Definition 10.1. Method (10.19) is called *stable*, if the generating polynomial $\varrho(\zeta)$ satisfies:

- i) The roots of $\varrho(\zeta)$ lie on or within the unit circle;
- ii) The multiplicity of the roots on the unit circle is at most two.

For the *order conditions* we introduce, similarly to formula (2.3), the linear difference operator

$$\begin{aligned} L(y, x, h) &= \varrho(E)y(x) - h^2 \sigma(E)y''(x) \\ &= \sum_{i=0}^k \left(\alpha_i y(x + ih) - h^2 \beta_i y''(x + ih) \right), \end{aligned} \quad (10.23)$$

where E is the shift operator. As in Definition 2.3 we now have:

Definition 10.2. Method (10.19) is *consistent of order p* if for all sufficiently smooth functions $y(x)$,

$$L(y, x, h) = \mathcal{O}(h^{p+2}). \quad (10.24)$$

The following theorem is then proved similarly to Theorem 2.4.

Theorem 10.3. *The multistep method (10.19) is of order p if and only if the following equivalent conditions hold:*

- i) $\sum_{i=0}^k \alpha_i = 0, \quad \sum_{i=0}^k i\alpha_i = 0$
 and $\sum_{i=0}^k \alpha_i i^q = q(q-1) \sum_{i=0}^k \beta_i i^{q-2}$ for $q = 2, \dots, p+1$,
- ii) $\varrho(e^h) - h^2 \sigma(e^h) = \mathcal{O}(h^{p+2})$ for $h \rightarrow 0$,
- iii) $\frac{\varrho(\zeta)}{(\log \zeta)^2} - \sigma(\zeta) = \mathcal{O}((\zeta - 1)^p)$ for $\zeta \rightarrow 1$.

□

As for Adams methods one easily verifies that the method (10.10) is of order k , and that (10.13) is of order $k+1$.

The following order barriers are similar to those of Theorems 3.5 and 3.9; their proofs are similar too (see, e.g., Dahlquist 1959, Henrici 1962):

Theorem 10.4. *The order p of a stable linear multistep method (10.19) satisfies*

$$\begin{aligned} p &\leq k+2 \quad \text{if } k \text{ is even,} \\ p &\leq k+1 \quad \text{if } k \text{ is odd.} \end{aligned}$$

□

Theorem 10.5. *Stable multistep methods (10.19) of order $k+2$ are symmetric, i.e.,*

$$\alpha_j = \alpha_{k-j}, \quad \beta_j = \beta_{k-j} \quad \text{for all } j.$$

□

Convergence

Theorem 10.6. *Suppose that method (10.19) is stable, of order p , and that the starting values satisfy*

$$y(x_j) - y_j = \mathcal{O}(h^{p+1}) \quad \text{for } j = 0, 1, \dots, k-1. \quad (10.25)$$

Then we have convergence of order p , i.e.,

$$\|y(x_n) - y_n\| \leq Ch^p \quad \text{for } 0 \leq hn \leq \text{Const.}$$

Proof. It is possible to develop a theory analogous to that of Sections III.2 - III.4. This is due to Dahlquist (1959) and can also be found in the book of Henrici (1962). We prefer to rewrite (10.19) in a one-step formulation of the form (8.4) and to apply directly the results of Section III.8 and III.9 (see Example 8.6). In order to achieve this goal, we could put $u_n = (y_{n+k-1}, \dots, y_n)^T$, which seems to be a natural choice. But then the corresponding matrix S does not satisfy the stability condition of Definition 8.8 because of the double roots of modulus 1. To overcome this difficulty we separate these roots. We split the characteristic polynomial $\varrho(\zeta)$ into

$$\varrho(\zeta) = \varrho_1(\zeta) \cdot \varrho_2(\zeta) \quad (10.26)$$

such that each polynomial ($l + k = m$)

$$\varrho_1(\zeta) = \sum_{i=0}^l \gamma_i \zeta^i, \quad \varrho_2(\zeta) = \sum_{i=0}^m \kappa_i \zeta^i \quad (10.27)$$

has only simple roots of modulus 1. Without loss of generality we assume in the sequel that $m \geq l$ and $\alpha_k = \gamma_l = \kappa_m = 1$. Using the shift operator E , method (10.19) can be written as

$$\varrho(E)y_n = h^2 \sigma(E)f_n.$$

The main idea is to introduce $\varrho_2(E)y_n$ as a new variable, say $h v_n$, so that the multistep formula becomes equivalent to the system

$$\varrho_1(E)v_n = h \sigma(E)f_n, \quad \varrho_2(E)y_n = h v_n. \quad (10.28)$$

Introducing the vector

$$u_n = (v_{n+l-1}, \dots, v_n, y_{n+m-1}, \dots, y_n)^T$$

formula (10.28) can be written as

$$u_{n+1} = S u_n + h \Phi(x_n, u_n, h) \quad (10.29a)$$

where

$$S = \begin{pmatrix} G & 0 \\ 0 & K \end{pmatrix}, \quad \Phi(x_n, u_n, h) = \begin{pmatrix} e_1 \psi(x_n, u_n, h) \\ e_1 v_n \end{pmatrix}. \quad (10.30)$$

The matrices G and K are the companion matrices

$$G = \begin{pmatrix} -\gamma_{l-1} & -\gamma_{l-2} & \cdots & \cdot & -\gamma_0 \\ 1 & 0 & \cdots & \cdot & 0 \\ & 1 & & \cdot & 0 \\ & & \ddots & \ddots & \vdots \\ & & & 1 & 0 \end{pmatrix}, \quad K = \begin{pmatrix} -\kappa_{m-1} & -\kappa_{m-2} & \cdots & \cdot & -\kappa_0 \\ 1 & 0 & \cdots & \cdot & 0 \\ & 1 & & \cdot & 0 \\ & & \ddots & \ddots & \vdots \\ & & & 1 & 0 \end{pmatrix},$$

$e_1 = (1, 0, \dots, 0)^T$, and $\psi = \psi(x_n, u_n, h)$ is implicitly defined by

$$\psi = \sum_{j=0}^{k-1} \beta_j f(x_n + jh, y_{n+j}) + \beta_k f(x_n + kh, h^2 \psi - \sum_{j=0}^{k-1} \alpha_j y_{n+j}). \quad (10.31)$$

In this formula ψ is written as a function of $x_n, (y_{n+k-1}, \dots, y_n)$ and h . But the second relation of (10.28) shows that each value $y_{n+k-1}, \dots, y_{n+m}$ can be expressed as a linear combination of the elements of u_n . Therefore ψ is in fact a function of (x_n, u_n, h) .

Formula (10.29a) defines our forward step procedure. The corresponding starting procedure is

$$\varphi(h) = (v_{l-1}, \dots, v_0, y_{m-1}, \dots, y_0)^T \quad (10.29b)$$

which, by (10.28), is uniquely determined by $(y_{k-1}, \dots, y_0)^T$. As correct value function we have

$$z(x, h) = \left(\frac{1}{h} \varrho_2(E) y(x + (l-1)h), \dots, \frac{1}{h} \varrho_2(E) y(x), y(x + (m-1)h), \dots, y(x) \right)^T. \quad (10.29c)$$

By our choice of $\varrho_1(\zeta)$ and $\varrho_2(\zeta)$ (both have only simple roots of modulus 1) the matrices G and K are power bounded. Therefore S is also power bounded and method (10.29) is *stable* in the sense of Definition 8.8.

We now verify the conditions of Definition 8.10 and for this start with the error in the initial values

$$d_0 = z(x_0, h) - \varphi(h).$$

The first l components of this vector are

$$\frac{1}{h} \varrho_2(E) y(x_j) - v_j = \frac{1}{h} \sum_{i=0}^m \kappa_i (y(x_{i+j}) - y_{i+j}), \quad j = 0, \dots, l-1$$

and the last m components are just

$$y(x_j) - y_j, \quad j = 0, \dots, m-1.$$

Thus hypothesis (10.25) ensures that $d_0 = \mathcal{O}(h^p)$. Consider next the local error at x_n ,

$$d_{n+1} = z(x_n + h, h) - S z(x_n, h) - h \Phi(x_n, z(x_n, h), h).$$

All components of d_{n+1} vanish except the first, which equals

$$d_{n+1}^{(1)} = \frac{1}{h} \varrho(E) y(x_n) - h \psi(x_n, z(x_n, h), h).$$

Using formula (10.31), an application of the mean value theorem yields

$$d_{n+1}^{(1)} = \frac{1}{h} L(y, x_n, h) + h^2 \beta_k f'(x_{n+k}, \eta) \cdot d_{n+1}^{(1)} \quad (10.32)$$

with η as in Lemma 2.2. We therefore have

$$d_{n+1} = \mathcal{O}(h^{p+1}) \quad \text{since} \quad L(y, x_n, h) = \mathcal{O}(h^{p+2}).$$

Finally Theorem 8.13 yields the stated convergence result. \square

Asymptotic Formula for the Global Error

Assume that the method (10.19) is stable and consistent of order p . The local truncation error of (10.29) is then given by

$$d_{n+1} = e_1 h^{p+1} C_{p+2} y^{(p+2)}(x_n) + \mathcal{O}(h^{p+2}) \quad (10.33)$$

with

$$C_{p+2} = \frac{1}{(p+2)!} \sum_{i=0}^k \left(\alpha_i i^{p+2} - (p+2)(p+1) \beta_i i^p \right).$$

Formula (10.33) can be verified by developing $L(y, x_n, h)$ into a Taylor series in (10.32). An application of Theorem 9.1 (if 1 is the only root of modulus 1 of $\varrho(\zeta)$) or of Theorem 9.2 shows that the global error of method (10.29) is of the form

$$u_h(x) - z(x, h) = e(x)h^p + \mathcal{O}(h^{p+1})$$

where $e(x)$ is the solution of

$$e'(x) = E \frac{\partial \Phi}{\partial u}(x, z(x, 0), 0) e(x) - E e_1 \cdot C_{p+2} y^{(p+2)}(x). \quad (10.34)$$

Here E is the matrix defined in (8.12). Since no h^p -term is present in the local error (10.33), it follows from (9.16) that $e(x) = Ee(x)$. Therefore (see Exercise 4a) this function can be written as

$$e(x) = \begin{pmatrix} \gamma(x) \mathbb{I} \\ \kappa(x) \mathbb{I} \end{pmatrix}.$$

A straightforward calculation of $\frac{\partial \Phi}{\partial u}(x, z(x, 0), 0)$ and Ee_1 (for details see Exercise 4) shows that (10.34) becomes equivalent to the system

$$\gamma'(x) = \frac{\sigma(1)}{\varrho'_1(1)} \frac{\partial f}{\partial y}(x, y(x)) \kappa(x) - \frac{C_{p+2}}{\varrho'_1(1)} y^{(p+2)}(x) \quad (10.35a)$$

$$\kappa'(x) = \frac{1}{\varrho'_2(1)} \gamma(x). \quad (10.35b)$$

Differentiating (10.35b) and inserting $\gamma'(x)$ from (10.35a), we finally obtain

$$\kappa''(x) = \frac{\partial f}{\partial y}(x, y(x)) \kappa(x) - C y^{(p+2)}(x) \quad (10.36)$$

with

$$C = \frac{C_{p+2}}{\sigma(1)}. \quad (10.37)$$

Here we have used the relation $\sigma(1) = \varrho'_1(1) \cdot \varrho'_2(1)$, which is an immediate consequence of (10.26), and the assumption that the order of the method is at least 1. The constant C in (10.37) is called the *error constant* of method (10.19). It plays the same role as (2.13) for first order equations.

Since the last component of the vector u_n is y_n we have the desired result

$$y_n - y(x_n) = \kappa(x_n)h^p + \mathcal{O}(h^{p+1})$$

with $\kappa(x)$ satisfying (10.36). Further terms in the asymptotic expansion of the global error can also be obtained by specializing the results of III.9.

Rounding Errors

A *direct* implementation of Störmer's methods, for which (10.19) specializes to

$$y_{n+1} - 2y_n + y_{n-1} = h^2 \sum_{i=0}^k \beta_i f_{n+i-k+1}, \quad (10.38)$$

by storing the y -values y_0, y_1, \dots, y_{k-1} and computing successively the values y_k, y_{k+1}, \dots with the help of (10.38) leads to numerical instabilities for small h . This instability is caused by the double root of $\varrho(\zeta)$ on the unit circle. It can be observed numerically in Fig. 10.3, where the left picture is a zoom of Fig. 10.2, while the right image contains the results of a code implementing (10.38) directly.

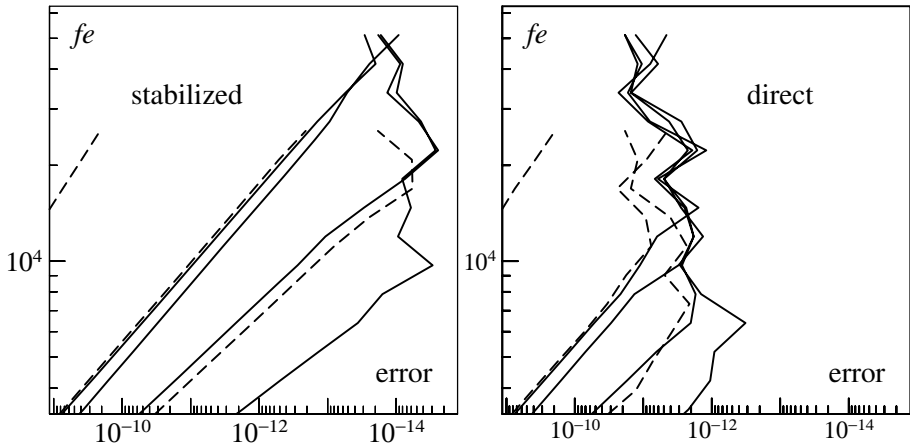


Fig. 10.3. Rounding errors caused by a direct application of (10.38)

In order to obtain the stabilized version of the algorithm, we apply the following two ideas:

- a) Split, as in (10.26), the polynomial $\varrho(\zeta)$ as $(\zeta - 1)(\zeta - 1)$. Then (10.28) leads to $h v_n = y_{n+1} - y_n$ and (10.38) becomes the mathematically equivalent formulation

$$v_n - v_{n-1} = h \sum_{i=0}^k \beta_i f_{n+i-k+1}, \quad y_{n+1} - y_n = h v_n. \quad (10.38')$$

Here the corresponding matrix S of (10.30) is stable.

- b) Avoid the use of $v_n = (y_{n+1} - y_n)/h$ for the computation of the starting values v_0, v_1, \dots, v_{k-2} , since the difference is a numerically unstable operation. Instead, add up the increments of the Runge-Kutta method, which you use for the computation of the starting values, directly.

These two ideas together then produce the “stabilized” results in Fig. 10.3 and Fig. 10.2.

Exercises

1. Compute the solution of Störmer’s problem (10.18) with one of the methods of this section.
2. a) Show that the generating functions of the coefficients σ_i and σ_j^* (defined in (10.11) and (10.14))

$$S(t) = \sum_{j=0}^{\infty} \sigma_j t^j, \quad S^*(t) = \sum_{j=0}^{\infty} \sigma_j^* t^j$$

satisfy

$$S(t) = \left(\frac{t}{\log(1-t)} \right)^2 \frac{1}{1-t}, \quad S^*(t) = \left(\frac{t}{\log(1-t)} \right)^2.$$

- b) Compute the coefficients d_j of

$$\sum_{j=0}^{\infty} d_j t^j = \left(\frac{\log(1-t)}{t} \right)^2 = \left(1 + \frac{t}{2} + \frac{t^2}{3} + \frac{t^3}{4} + \dots \right)^2$$

and derive a recurrence relation for the σ_j and σ_j^* .

- c) Prove that $\sigma_j^* = \sigma_j - \sigma_{j-1}$.
3. Let $\varrho(\zeta)$ be a polynomial of degree k which has 1 as root of multiplicity 2. Then there exists a unique $\sigma(\zeta)$ such that the corresponding method is of order $k+1$.
 4. Consider the method (10.29) and, for simplicity, assume the differential equation to be a scalar one.
 - a) For any vector w in \mathbb{R}^k the image vector Ew , with E given by (8.12), satisfies

$$Ew = \begin{pmatrix} \gamma \mathbb{1} \\ \kappa \mathbb{1} \end{pmatrix}$$

where γ, κ are real numbers and $\mathbb{1}$ is the vector with all elements equal to 1. The dimensions of $\gamma \mathbb{1}$ and $\kappa \mathbb{1}$ are l and m , respectively.

- b) Verify that for $e_1 = (1, 0, \dots, 0)^T$,

$$E \begin{pmatrix} \alpha e_1 \\ \beta e_1 \end{pmatrix} = \begin{pmatrix} (\alpha/\varrho'_1(1))\mathbb{I} \\ (\beta/\varrho'_2(1))\mathbb{I} \end{pmatrix}.$$

- c) Show that

$$E \frac{\partial \Phi}{\partial u}(x, z(x, 0), 0) \begin{pmatrix} \gamma \mathbb{I} \\ \kappa \mathbb{I} \end{pmatrix} = \begin{pmatrix} (\sigma(1)/\varrho'_1(1))(\partial f/\partial y)(x, y(x))\kappa \mathbb{I} \\ (1/\varrho'_2(1))\gamma \mathbb{I} \end{pmatrix}.$$

Hint. With $Y_n = (y_{n+k-1}, \dots, y_n)^T$ the formula (10.31) expresses ψ as a function of (x_n, Y_n, h) . The second formula of (10.28) relates Y_n and u_n as

$$KY_n = Lu_n + \mathcal{O}(h) \quad \text{where} \quad K\mathbb{I} = L \begin{pmatrix} 0 \\ \mathbb{I} \end{pmatrix}$$

and K is invertible. Use the chain rule for the computation of $\partial\psi/\partial u$. See also Exercise 2 of Section III.4 and Exercise 1 of Section III.9.

5. Compute the error constant (10.37) for the methods (10.10) and (10.13).

Result. σ_k and σ_{k+1}^* , respectively.

Appendix. Fortran Codes

... but the software is in various states of development from experimental (a euphemism for badly written) to what we might call ...
(C.W. Gear, in Aiken 1985)

Several Fortran codes have been developed for our numerical computations. Those of the first edition have been improved and several new options have been included, e.g., automatic choice of initial step size, stiffness detection, dense output. We have seen many of the ideas, which are incorporated in these codes, in the programs of P. Deuflhard, A.C. Hindmarsh and L.F. Shampine.

Experiences with all of our codes are welcome. The programs can be obtained from the authors' homepage (<http://www.unige.ch/~hairer>).

Address: Section de Mathématiques, Case postale 240,

CH-1211 Genève 24, Switzerland

E-mail: Ernst.Hairer@math.unige.ch Gerhard.Wanner@math.unige.ch

Driver for the Code DOPRI5

The driver given here is for the differential equation (II.0.1) with initial values and x_{end} given in (II.0.2). This is the problem AREN of Section II.10. The subroutine FAREN ("F for AREN") computes the right-hand side of this differential equation. The subroutine SOLOUT ("Solution out"), which is called by DOPRI5 after every successful step, and the dense output routine CONTD5 are used to print the solution at equidistant points. The (optional) common block STATD5 gives statistical information after the call to DOPRI5. The common blocks COD5R and COD5I transfer the necessary information to CONTD5.

```
      IMPLICIT REAL*8 (A-H,O-Z)
      PARAMETER (NDGL=4,LWORK=8*NDGL+10,LIWORK=10)
      PARAMETER (NRDENS=2,LRCONT=5*NRDENS+2,LICONT=NRDENS+1)
      DIMENSION Y(NDGL),WORK(LWORK),IWORK(LIWORK)
      COMMON/STATD5/NFCN,NSTEP,NACCPT,NREJCT
      COMMON /COD5R/RCONT(LRCONT)
      COMMON /COD5I/ICONT(LICONT)
      EXTERNAL FAREN,SOLOUT
C --- DIMENSION OF THE SYSTEM
      N=NDGL
C --- OUTPUT ROUTINE (AND DENSE OUTPUT) IS USED DURING INTEGRATION
      IOUT=2
```

```

C --- INITIAL VALUES AND ENDPOINT OF INTEGRATION
      X=0.0D0
      Y(1)=0.994D0
      Y(2)=0.0D0
      Y(3)=0.0D0
      Y(4)=-2.00158510637908252240537862224D0
      XEND=17.0652165601579625588917206249D0
C --- REQUIRED (RELATIVE AND ABSOLUTE) TOLERANCE
      ITOL=0
      RTOL=1.0D-7
      ATOL=RTOL
C --- DEFAULT VALUES FOR PARAMETERS
      DO 10 I=1,10
        IWORK(I)=0
      10  WORK(I)=0.0D0
C --- DENSE OUTPUT IS USED FOR THE TWO POSITION COORDINATES 1 AND 2
      IWORK(5)=NRDENS
      ICONT(2)=1
      ICONT(3)=2
C --- CALL OF THE SUBROUTINE DOPRI5
      CALL DOPRI5(N,FAREN,X,Y,XEND,
+              RTOL,ATOL,ITOL,
+              SOLOUT,IOUT,
+              WORK,LWORK,IWORK,LIWORK,LRCNT,LICONT,IDID)
C --- PRINT FINAL SOLUTION
      WRITE (6,99) Y(1),Y(2)
      99  FORMAT(1X,'X = XEND      Y =',2E18.10)
C --- PRINT STATISTICS
      WRITE (6,91) RTOL,NFCN,NSTEP,NACCPY,NREJCT
      91  FORMAT('      tol=',D8.2,'      fcn=',I5,' step=',I4,
+              '      acct=',I4,'      rejct=',I3)
      STOP
      END
C
      SUBROUTINE SOLOUT (NR,XOLD,X,Y,N,IRTRN)
C --- PRINTS SOLUTION AT EQUIDISTANT OUTPUT-POINTS BY USING "CONTD5"
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION Y(N)
      COMMON /INTERN/XOUT
      IF (NR.EQ.1) THEN
        WRITE (6,99) X,Y(1),Y(2),NR-1
        XOUT=X+2.0D0
      ELSE
      10  CONTINUE
        IF (X.GE.XOUT) THEN
          WRITE (6,99) XOUT,CONTD5(1,XOUT),CONTD5(2,XOUT),NR-1
          XOUT=XOUT+2.0D0
          GOTO 10
        END IF
      END IF
      99  FORMAT(1X,'X =',F6.2,'      Y =',2E18.10,'      NSTEP =',I4)
      RETURN
      END
C
      SUBROUTINE FAREN(N,X,Y,F)
C --- ARENSTORF ORBIT
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION Y(N),F(N)
      AMU=0.012277471D0
      AMUP=1.0D0-AMU

```

```

F(1)=Y(3)
F(2)=Y(4)
R1=(Y(1)+AMU)**2+Y(2)**2
R1=R1*SQRT(R1)
R2=(Y(1)-AMU)**2+Y(2)**2
R2=R2*SQRT(R2)
F(3)=Y(1)+2*Y(4)-AMUP*(Y(1)+AMU)/R1-AMU*(Y(1)-AMU)/R2
F(4)=Y(2)-2*Y(3)-AMUP*Y(2)/R1-AMU*Y(2)/R2
RETURN
END

```

The result, obtained on an Apollo workstation, is the following:

```

X = 0.00    Y = 0.9940000000E+00    0.0000000000E+00    NSTEP = 0
X = 2.00    Y = -0.5798781411E+00    0.6090775251E+00    NSTEP = 60
X = 4.00    Y = -0.1983335270E+00    0.1137638086E+01    NSTEP = 73
X = 6.00    Y = -0.4735743943E+00    0.2239068118E+00    NSTEP = 91
X = 8.00    Y = -0.1174553350E+01    -0.2759466982E+00    NSTEP = 110
X = 10.00   Y = -0.8398073466E+00    0.4468302268E+00    NSTEP = 122
X = 12.00   Y = 0.1314712468E-01    -0.8385751499E+00    NSTEP = 145
X = 14.00   Y = -0.6031129504E+00    -0.9912598031E+00    NSTEP = 159
X = 16.00   Y = 0.2427110999E+00    -0.3899948833E+00    NSTEP = 177
X = XEND    Y = 0.9940021016E+00    0.8911185978E-05
      tol=0.10E-06    fcn= 1442 step= 240 acpt= 216 reject= 22

```

Subroutine DOPRI5

Explicit Runge-Kutta code based on the method of Dormand & Prince (see Table 5.2 of Section II.5). It is provided with the step control algorithm of Section II.4 and the dense output of Section II.6.

```

      SUBROUTINE DOPRI5(N,FCN,X,Y,XEND,
+                      RTOL,ATOL,ITOL,
+                      SOLOUT,IOUT,
+                      WORK,LWORK,IWORK,LIWORK,LRCONT,LICONT,IDID)
C -----
C      NUMERICAL SOLUTION OF A SYSTEM OF FIRST ORDER
C      ORDINARY DIFFERENTIAL EQUATIONS Y'=F(X,Y).
C      THIS IS AN EXPLICIT RUNGE-KUTTA METHOD OF ORDER (4)5
C      DUE TO DORMAND & PRINCE (WITH STEPSIZE CONTROL AND
C      DENSE OUTPUT).
C
C      AUTHORS: E. HAIRER AND G. WANNER
C               UNIVERSITE DE GENEVE, DEPT. DE MATHEMATIQUES
C               CH-1211 GENEVE 24, SWITZERLAND
C               E-MAIL: HAIRER@UNI2A.UNIGE.CH, WANNER@UNI2A.UNIGE.CH
C
C      THIS CODE IS DESCRIBED IN:
C      E. HAIRER, S.P. NORSETT AND G. WANNER, SOLVING ORDINARY
C      DIFFERENTIAL EQUATIONS I. NONSTIFF PROBLEMS. 2ND EDITION.
C      SPRINGER SERIES IN COMPUTATIONAL MATHEMATICS,
C      SPRINGER-VERLAG (1993)
C

```

```

C      VERSION OF OCTOBER 3, 1991
C
C      INPUT PARAMETERS
C      -----
C      N          DIMENSION OF THE SYSTEM
C
C      FCN        NAME (EXTERNAL) OF SUBROUTINE COMPUTING THE
C                VALUE OF F(X,Y):
C                SUBROUTINE FCN(N,X,Y,F)
C                REAL*8 X,Y(N),F(N)
C                F(1)=...      ETC.
C
C      X          INITIAL X-VALUE
C
C      Y(N)       INITIAL VALUES FOR Y
C
C      XEND       FINAL X-VALUE (XEND-X MAY BE POSITIVE OR NEGATIVE)
C
C      RTOL,ATOL  RELATIVE AND ABSOLUTE ERROR TOLERANCES. THEY
C                CAN BE BOTH SCALARS OR ELSE BOTH VECTORS OF LENGTH N.
C
C      ITOL       SWITCH FOR RTOL AND ATOL:
C                ITOL=0: BOTH RTOL AND ATOL ARE SCALARS.
C                THE CODE KEEPS, ROUGHLY, THE LOCAL ERROR OF
C                Y(I) BELOW RTOL*ABS(Y(I))+ATOL
C                ITOL=1: BOTH RTOL AND ATOL ARE VECTORS.
C                THE CODE KEEPS THE LOCAL ERROR OF Y(I) BELOW
C                RTOL(I)*ABS(Y(I))+ATOL(I).
C
C      SOLOUT     NAME (EXTERNAL) OF SUBROUTINE PROVIDING THE
C                NUMERICAL SOLUTION DURING INTEGRATION.
C                IF IOUT.GE.1, IT IS CALLED AFTER EVERY SUCCESSFUL STEP.
C                SUPPLY A DUMMY SUBROUTINE IF IOUT=0.
C                IT MUST HAVE THE FORM
C                SUBROUTINE SOLOUT (NR,XOLD,X,Y,N,IRTRN)
C                REAL*8 X,Y(N)
C                ....
C                SOLOUT FURNISHES THE SOLUTION "Y" AT THE NR-TH
C                GRID-POINT "X" (THEREBY THE INITIAL VALUE IS
C                THE FIRST GRID-POINT).
C                "XOLD" IS THE PRECEEDING GRID-POINT.
C                "IRTRN" SERVES TO INTERRUPT THE INTEGRATION. IF IRTRN
C                IS SET <0, DOPRI5 WILL RETURN TO THE CALLING PROGRAM.
C
C      ----- CONTINUOUS OUTPUT: -----
C                DURING CALLS TO "SOLOUT", A CONTINUOUS SOLUTION
C                FOR THE INTERVAL [XOLD,X] IS AVAILABLE THROUGH
C                THE FUNCTION
C                >>> CONTD5(I,S) <<<
C                WHICH PROVIDES AN APPROXIMATION TO THE I-TH
C                COMPONENT OF THE SOLUTION AT THE POINT S. THE VALUE
C                S SHOULD LIE IN THE INTERVAL [XOLD,X].
C
C      IOUT       SWITCH FOR CALLING THE SUBROUTINE SOLOUT:
C                IOUT=0: SUBROUTINE IS NEVER CALLED
C                IOUT=1: SUBROUTINE IS USED FOR OUTPUT.
C                IOUT=2: DENSE OUTPUT IS PERFORMED IN SOLOUT
C                       (IN THIS CASE WORK(5) MUST BE SPECIFIED)
C
C      WORK       ARRAY OF WORKING SPACE OF LENGTH "LWORK".

```



```

C          "LWORK" MUST BE AT LEAST 8*N+10
C
C      LWORK      DECLARED LENGHT OF ARRAY "WORK".
C
C      IWORK      INTEGER WORKING SPACE OF LENGHT "LIWORK".
C                  IWORK(1),...,IWORK(5) SERVE AS PARAMETERS
C                  FOR THE CODE. FOR STANDARD USE, SET THEM
C                  TO ZERO BEFORE CALLING.
C                  "LIWORK" MUST BE AT LEAST 10 .
C
C      LIWORK     DECLARED LENGHT OF ARRAY "IWORK".
C
C      LRCONT     DECLARED LENGTH OF COMMON BLOCK
C                  >>> COMMON /COD5R/RCONT(LRCONT) <<<
C                  WHICH MUST BE DECLARED IN THE CALLING PROGRAM.
C                  "LRCONT" MUST BE AT LEAST
C                      5 * NRDENS + 2
C                  WHERE NRDENS=IWORK(5) (SEE BELOW).
C
C      LICONT     DECLARED LENGTH OF COMMON BLOCK
C                  >>> COMMON /COD5I/ICONT(LICONT) <<<
C                  WHICH MUST BE DECLARED IN THE CALLING PROGRAM.
C                  "LICONT" MUST BE AT LEAST
C                      NRDENS + 1
C                  THESE COMMON BLOCKS ARE USED FOR STORING THE COEFFICIENTS
C                  OF THE CONTINUOUS SOLUTION AND MAKES THE CALLING LIST FOR
C                  THE FUNCTION "CONTD5" AS SIMPLE AS POSSIBLE.
C
C-----
C
C      SOPHISTICATED SETTING OF PARAMETERS
C      -----
C          SEVERAL PARAMETERS (WORK(1),...,IWORK(1),...) ALLOW
C          TO ADAPT THE CODE TO THE PROBLEM AND TO THE NEEDS OF
C          THE USER. FOR ZERO INPUT, THE CODE CHOOSES DEFAULT VALUES.
C
C      WORK(1)    UROUND, THE ROUNDING UNIT, DEFAULT 2.3D-16.
C
C      WORK(2)    THE SAFETY FACTOR IN STEP SIZE PREDICTION,
C                  DEFAULT 0.9D0.
C
C      WORK(3), WORK(4)  PARAMETERS FOR STEP SIZE SELECTION
C                  THE NEW STEP SIZE IS CHOSEN SUBJECT TO THE RESTRICTION
C                      WORK(3) <= HNEW/HOLD <= WORK(4)
C                  DEFAULT VALUES: WORK(3)=0.2D0, WORK(4)=10.D0
C
C      WORK(5)    IS THE "BETA" FOR STABILIZED STEP SIZE CONTROL
C                  (SEE SECTION IV.2). LARGER VALUES OF BETA ( <= 0.1 )
C                  MAKE THE STEP SIZE CONTROL MORE STABLE. DOPRI5 NEEDS
C                  A LARGER BETA THAN HIGHAM & HALL. NEGATIVE WORK(5)
C                  PROVOKE BETA=0.
C                  DEFAULT 0.04D0.
C
C      WORK(6)    MAXIMAL STEP SIZE, DEFAULT XEND-X.
C
C      WORK(7)    INITIAL STEP SIZE, FOR WORK(7)=0.D0 AN INITIAL GUESS
C                  IS COMPUTED WITH HELP OF THE FUNCTION HINIT
C
C      IWORK(1)   THIS IS THE MAXIMAL NUMBER OF ALLOWED STEPS.
C                  THE DEFAULT VALUE (FOR IWORK(1)=0) IS 100000.

```

```

C
C      IWORK(2) SWITCH FOR THE CHOICE OF THE COEFFICIENTS
C              IF IWORK(2).EQ.1 METHOD DOPRI5 OF DORMAND AND PRINCE
C              (TABLE 5.2 OF SECTION II.5).
C              AT THE MOMENT THIS IS THE ONLY POSSIBLE CHOICE.
C              THE DEFAULT VALUE (FOR IWORK(2)=0) IS IWORK(2)=1.
C
C
C      IWORK(3) SWITCH FOR PRINTING ERROR MESSAGES
C              IF IWORK(3).LT.0 NO MESSAGES ARE BEING PRINTED
C              IF IWORK(3).GT.0 MESSAGES ARE PRINTED WITH
C              WRITE (IWORK(3),*) ...
C              DEFAULT VALUE (FOR IWORK(3)=0) IS IWORK(3)=6
C
C
C      IWORK(4) TEST FOR STIFFNESS IS ACTIVATED AFTER STEP NUMBER
C              J*IWORK(4) (J INTEGER), PROVIDED IWORK(4).GT.0.
C              FOR NEGATIVE IWORK(4) THE STIFFNESS TEST IS
C              NEVER ACTIVATED; DEFAULT VALUE IS IWORK(4)=1000
C
C
C      IWORK(5) = NRDENS = NUMBER OF COMPONENTS, FOR WHICH DENSE OUTPUT
C              IS REQUIRED; DEFAULT VALUE IS IWORK(5)=0;
C              FOR 0 < NRDENS < N THE COMPONENTS (FOR WHICH DENSE
C              OUTPUT IS REQUIRED) HAVE TO BE SPECIFIED IN
C              ICONT(2),...,ICONT(NRDENS+1);
C              FOR NRDENS=N THIS IS DONE BY THE CODE.
C
C-----
C
C      OUTPUT PARAMETERS
C      -----
C
C      X          X-VALUE FOR WHICH THE SOLUTION HAS BEEN COMPUTED
C                  (AFTER SUCCESSFUL RETURN X=XEND).
C
C
C      Y(N)       NUMERICAL SOLUTION AT X
C
C
C      H          PREDICTED STEP SIZE OF THE LAST ACCEPTED STEP
C
C
C      IDID       REPORTS ON SUCCESSFULNESS UPON RETURN:
C                  IDID= 1 COMPUTATION SUCCESSFUL,
C                  IDID= 2 COMPUT. SUCCESSFUL (INTERRUPTED BY SOLOUT)
C                  IDID=-1 INPUT IS NOT CONSISTENT,
C                  IDID=-2 LARGER NMAX IS NEEDED,
C                  IDID=-3 STEP SIZE BECOMES TOO SMALL.
C                  IDID=-4 PROBLEM IS PROBABLY STIFF (INTERRUPTED).
C
C-----
C *** **
C
C      DECLARATIONS
C *** **
C      IMPLICIT REAL*8 (A-H,O-Z)
C      DIMENSION Y(N),ATOL(1),RTOL(1),WORK(LWORK),IWORK(LIWORK)
C      LOGICAL ARRET
C      EXTERNAL FCN,SOLOUT
C      COMMON/STATD5/NFCN,NSTEP,NACCPT,NREJCT
C --- COMMON STATD5 CAN BE INSPECTED FOR STATISTICAL PURPOSES:
C ---   NFCN      NUMBER OF FUNCTION EVALUATIONS
C ---   NSTEP     NUMBER OF COMPUTED STEPS
C ---   NACCPT    NUMBER OF ACCEPTED STEPS
C ---   NREJCT    NUMBER OF REJECTED STEPS (AFTER AT LEAST ONE STEP
C                HAS BEEN ACCEPTED)
C
C      .....

```

Subroutine DOP853

Explicit Runge-Kutta code of order 8 based on the method of Dormand & Prince, described in Section II.5. The local error estimation and the step size control is based on embedded formulas of orders 5 and 3 (see Section II.10). This method is provided with a dense output of order 7. In the following description we have omitted the parts which are identical to those for DOPRI5.

```

      SUBROUTINE DOP853(N,FCN,X,Y,XEND,
+          RTOL,ATOL,ITOL,
+          SOLOUT,IOUT,
+          WORK,LWORK,IWORK,LIWORK,LRCONT,LICONT,IDID)
C -----
C   NUMERICAL SOLUTION OF A SYSTEM OF FIRST ORDER
C   ORDINARY DIFFERENTIAL EQUATIONS  Y'=F(X,Y).
C   THIS IS AN EXPLICIT RUNGE-KUTTA METHOD OF ORDER 8(5,3)
C   DUE TO DORMAND & PRINCE (WITH STEPSIZE CONTROL AND
C   DENSE OUTPUT)
C   .....
C   VERSION OF NOVEMBER 29, 1992
C   .....
C   -----  CONTINUOUS OUTPUT: -----
C   DURING CALLS TO "SOLOUT", A CONTINUOUS SOLUTION
C   FOR THE INTERVAL [XOLD,X] IS AVAILABLE THROUGH
C   THE FUNCTION
C       >>>  CONTD8(I,S)  <<<
C   WHICH PROVIDES AN APPROXIMATION TO THE I-TH
C   .....
C   WORK      ARRAY OF WORKING SPACE OF LENGTH "LWORK".
C   "LWORK" MUST BE AT LEAST  11*N+10
C   .....
C   LRCONT    DECLARED LENGTH OF COMMON BLOCK
C   >>>  COMMON /COD8R/RCONT(LRCONT)  <<<
C   WHICH MUST BE DECLARED IN THE CALLING PROGRAM.
C   "LRCONT" MUST BE AT LEAST
C       8 * NRDENS + 2
C   WHERE NRDENS=IWORK(5) (SEE BELOW).
C   .....
C   LICONT    DECLARED LENGTH OF COMMON BLOCK
C   >>>  COMMON /COD8I/ICONT(LICONT)  <<<
C   WHICH MUST BE DECLARED IN THE CALLING PROGRAM.
C   "LICONT" MUST BE AT LEAST
C       NRDENS + 1
C   THESE COMMON BLOCKS ARE USED FOR STORING THE COEFFICIENTS
C   OF THE CONTINUOUS SOLUTION AND MAKES THE CALLING LIST FOR
C   THE FUNCTION "CONTD8" AS SIMPLE AS POSSIBLE.
C   .....
C   WORK(3), WORK(4)  PARAMETERS FOR STEP SIZE SELECTION
C   THE NEW STEP SIZE IS CHOSEN SUBJECT TO THE RESTRICTION
C   WORK(3) <= HNEW/HOLD <= WORK(4)
C   DEFAULT VALUES: WORK(3)=0.333D0, WORK(4)=6.D0
C   .....

```

Subroutine ODEX

Extrapolation code for $y' = f(x, y)$, based on the GBS algorithm (Section II.9). It uses variable order and variable step sizes and is provided with a high-order dense output. Again, the missing parts in the description are identical to those of DOPRI5.

```

      SUBROUTINE ODEX(N,FCN,X,Y,XEND,H,
+          RTOL,ATOL,ITOL,
+          SOLOUT,IOUT,
+          WORK,LWORK,IWORK,LIWORK,LRCONT,LICONT,IDID)
C -----
C   NUMERICAL SOLUTION OF A SYSTEM OF FIRST ORDER
C   ORDINARY DIFFERENTIAL EQUATIONS  Y'=F(X,Y).
C   THIS IS AN EXTRAPOLATION-ALGORITHM (GBS), BASED ON THE
C   EXPLICIT MIDPOINT RULE (WITH STEPSIZE CONTROL,
C   ORDER SELECTION AND DENSE OUTPUT).
C
C   AUTHORS: E. HAIRER AND G. WANNER
C            UNIVERSITE DE GENEVE, DEPT. DE MATHEMATIQUES
C            CH-1211 GENEVE 24, SWITZERLAND
C            E-MAIL: HAIRER@UNI2A.UNIGE.CH, WANNER@UNI2A.UNIGE.CH
C            DENSE OUTPUT WRITTEN BY E. HAIRER AND A. OSTERMANN
C
C   .....
C   VERSION DECEMBER 18, 1991
C   .....
C
C   H          INITIAL STEP SIZE GUESS;
C              H=1.DO/(NORM OF F'), USUALLY 1.D-1 OR 1.D-3, IS GOOD.
C              THIS CHOICE IS NOT VERY IMPORTANT, THE CODE QUICKLY
C              ADAPTS ITS STEP SIZE. WHEN YOU ARE NOT SURE, THEN
C              STUDY THE CHOSEN VALUES FOR A FEW
C              STEPS IN SUBROUTINE "SOLOUT".
C              (IF H=0.DO, THE CODE PUTS H=1.D-4).
C
C   .....
C
C   ----- CONTINUOUS OUTPUT (IF IOUT=2): -----
C   DURING CALLS TO "SOLOUT", A CONTINUOUS SOLUTION
C   FOR THE INTERVAL [XOLD,X] IS AVAILABLE THROUGH
C   THE REAL*8 FUNCTION
C              >>> CONTEX(I,S) <<<
C   WHICH PROVIDES AN APPROXIMATION TO THE I-TH
C   COMPONENT OF THE SOLUTION AT THE POINT S. THE VALUE
C   S SHOULD LIE IN THE INTERVAL [XOLD,X].
C
C   .....
C
C   WORK       ARRAY OF WORKING SPACE OF LENGTH "LWORK".
C              SERVES AS WORKING SPACE FOR ALL VECTORS.
C              "LWORK" MUST BE AT LEAST
C              N*(KM+5)+5*KM+10+2*KM*(KM+1)*NRDENS
C              WHERE NRDENS=IWORK(8) (SEE BELOW) AND
C              KM=9          IF IWORK(2)=0
C              KM=IWORK(2)   IF IWORK(2).GT.0
C              WORK(1),...,WORK(10) SERVE AS PARAMETERS
C              FOR THE CODE. FOR STANDARD USE, SET THESE
C              PARAMETERS TO ZERO BEFORE CALLING.
C
C   .....

```

```

C
C      IWORK      INTEGER WORKING SPACE OF LENGTH "LIWORK".
C                  "LIWORK" MUST BE AT LEAST
C                      2*KM+10+NRDENS
C                  IWORK(1),...,IWORK(9) SERVE AS PARAMETERS
C                  FOR THE CODE. FOR STANDARD USE, SET THESE
C                  PARAMETERS TO ZERO BEFORE CALLING.
C
C      .....
C      LRCONT     DECLARED LENGTH OF COMMON BLOCK
C                  >>> COMMON /CONTR/RCONT(LRCONT) <<<
C                  WHICH MUST BE DECLARED IN THE CALLING PROGRAM.
C                  "LRCONT" MUST BE AT LEAST
C                      ( 2 * KM + 5 ) * NRDENS + 2
C                  WHERE KM=IWORK(2) AND NRDENS=IWORK(8) (SEE BELOW).
C
C      LICONT     DECLARED LENGTH OF COMMON BLOCK
C                  >>> COMMON /CONTI/ICONT(LICONT) <<<
C                  WHICH MUST BE DECLARED IN THE CALLING PROGRAM.
C                  "LICONT" MUST BE AT LEAST
C                      NRDENS + 2
C                  THESE COMMON BLOCKS ARE USED FOR STORING THE COEFFICIENTS
C                  OF THE CONTINUOUS SOLUTION AND MAKES THE CALLING LIST FOR
C                  THE FUNCTION "CONTEX" AS SIMPLE AS POSSIBLE.
C
C      .....
C      WORK(2)    MAXIMAL STEP SIZE, DEFAULT XEND-X.
C
C      WORK(3)    STEP SIZE IS REDUCED BY FACTOR WORK(3), IF THE
C                  STABILITY CHECK IS NEGATIVE, DEFAULT 0.5.
C
C      WORK(4), WORK(5)  PARAMETERS FOR STEP SIZE SELECTION
C                  THE NEW STEP SIZE FOR THE J-TH DIAGONAL ENTRY IS
C                  CHOSEN SUBJECT TO THE RESTRICTION
C                      FACMIN/WORK(5) <= HNEW(J)/HOLD <= 1/FACMIN
C                  WHERE FACMIN=WORK(4)**(1/(2*J-1))
C                  DEFAULT VALUES: WORK(4)=0.02D0, WORK(5)=4.D0
C
C      WORK(6), WORK(7)  PARAMETERS FOR THE ORDER SELECTION
C                  STEP SIZE IS DECREASED IF    W(K-1) <= W(K)*WORK(6)
C                  STEP SIZE IS INCREASED IF    W(K) <= W(K-1)*WORK(7)
C                  DEFAULT VALUES: WORK(6)=0.8D0, WORK(7)=0.9D0
C
C      WORK(8), WORK(9)  SAFETY FACTORS FOR STEP CONTROL ALGORITHM
C                  HNEW=H*WORK(9)*(WORK(8)*TOL/ERR)**(1/(J-1))
C                  DEFAULT VALUES: WORK(8)=0.65D0,
C                  WORK(9)=0.94D0 IF "HOPE FOR CONVERGENCE"
C                  WORK(9)=0.90D0 IF "NO HOPE FOR CONVERGENCE"
C
C      .....
C      IWORK(2)    THE MAXIMUM NUMBER OF COLUMNS IN THE EXTRAPOLATION
C                  TABLE. THE DEFAULT VALUE (FOR IWORK(2)=0) IS 9.
C                  IF IWORK(2).NE.0 THEN IWORK(2) SHOULD BE .GE.3.
C
C      IWORK(3)    SWITCH FOR THE STEP SIZE SEQUENCE (EVEN NUMBERS ONLY)
C                  IF IWORK(3).EQ.1 THEN 2,4,6,8,10,12,14,16,...
C                  IF IWORK(3).EQ.2 THEN 2,4,8,12,16,20,24,28,...
C                  IF IWORK(3).EQ.3 THEN 2,4,6,8,12,16,24,32,...
C                  IF IWORK(3).EQ.4 THEN 2,6,10,14,18,22,26,30,...
C                  IF IWORK(3).EQ.5 THEN 4,8,12,16,20,24,28,32,...

```

```

C          THE DEFAULT VALUE IS IWORK(3)=1 IF IOUT.LE.1;
C          THE DEFAULT VALUE IS IWORK(3)=4 IF IOUT.GE.2.
C
C  IWORK(4)  STABILITY CHECK IS ACTIVATED AT MOST IWORK(4) TIMES IN
C             ONE LINE OF THE EXTRAP. TABLE, DEFAULT IWORK(4)=1.
C
C  IWORK(5)  STABILITY CHECK IS ACTIVATED ONLY IN THE LINES
C             1 TO IWORK(5) OF THE EXTRAP. TABLE, DEFAULT IWORK(5)=1.
C
C  IWORK(6)  IF IWORK(6)=0 ERROR ESTIMATOR IN THE DENSE
C             OUTPUT FORMULA IS ACTIVATED. IT CAN BE SUPPRESSED
C             BY PUTTING IWORK(6)=1.
C             DEFAULT IWORK(6)=0 (IF IOUT.GE.2).
C
C  IWORK(7)  DETERMINES THE DEGREE OF INTERPOLATION FORMULA
C             MU = 2 * KAPPA - IWORK(7) + 1
C             IWORK(7) SHOULD LIE BETWEEN 1 AND 6
C             DEFAULT IWORK(7)=4 (IF IWORK(7)=0).
C
C  IWORK(8)  = NRDENS = NUMBER OF COMPONENTS, FOR WHICH DENSE OUTPUT
C             IS REQUIRED
C
C  IWORK(10),...,IWORK(NRDENS+9) INDICATE THE COMPONENTS, FOR WHICH
C             DENSE OUTPUT IS REQUIRED
C
C .....
C
C  IDID      REPORTS ON SUCCESSFULNESS UPON RETURN:
C             IDID=1  COMPUTATION SUCCESSFUL,
C             IDID=-1 COMPUTATION UNSUCCESSFUL.
C
C .....

```

Subroutine ODEX2

Extrapolation code for second order differential equations $y'' = f(x, y)$ (Section II.14). It uses variable order and variable step sizes and is provided with a high-order dense output. The missing parts of the description are identical to those of ODEX.

```

          SUBROUTINE ODEX2(N,FCN,X,Y,YP,XEND,H,
+                      RTOL,ATOL,ITOL,
+                      SOLOUT,IOUT,
+                      WORK,LWORK,IWORK,LIWORK,LRCONT,LICONT,IDID)
C -----
C  NUMERICAL SOLUTION OF A SYSTEM OF SECOND ORDER
C  ORDINARY DIFFERENTIAL EQUATIONS  Y''=F(X,Y).
C  THIS IS AN EXTRAPOLATION-ALGORITHM, BASED ON
C  THE STOERMER RULE (WITH STEPSIZE CONTROL
C  ORDER SELECTION AND DENSE OUTPUT).
C .....
C
C  VERSION MARCH 30, 1992
C .....
C
C  Y(N)      INITIAL VALUES FOR Y
C

```

```

C      YP(N)      INITIAL VALUES FOR Y'
C      .....
C
C      ITOL      SWITCH FOR RTOL AND ATOL:
C                ITOL=0:  BOTH RTOL AND ATOL ARE SCALARS.
C                THE CODE KEEPS, ROUGHLY, THE LOCAL ERROR OF
C                Y(I) BELOW RTOL*ABS(Y(I))+ATOL
C                YP(I) BELOW RTOL*ABS(YP(I))+ATOL
C                ITOL=1:  BOTH RTOL AND ATOL ARE VECTORS.
C                THE CODE KEEPS THE LOCAL ERROR OF
C                Y(I) BELOW RTOL(I)*ABS(Y(I))+ATOL(I).
C                YP(I) BELOW RTOL(I+N)*ABS(YP(I))+ATOL(I+N).
C
C      SOLOUT     NAME (EXTERNAL) OF SUBROUTINE PROVIDING THE
C                NUMERICAL SOLUTION DURING INTEGRATION.
C                IF IOUT>=1, IT IS CALLED AFTER EVERY SUCCESSFUL STEP.
C                SUPPLY A DUMMY SUBROUTINE IF IOUT=0.
C                IT MUST HAVE THE FORM
C                SUBROUTINE SOLOUT (NR,XOLD,X,Y,YP,N,IRTRN)
C                REAL*8 X,Y(N),YP(N)
C                ....
C                SOLOUT FURNISHES THE SOLUTIONS "Y, YP" AT THE NR-TH
C                GRID-POINT "X" (THEREBY THE INITIAL VALUE IS
C                THE FIRST GRID-POINT).
C                "XOLD" IS THE PRECEDING GRID-POINT.
C                "IRTRN" SERVES TO INTERRUPT THE INTEGRATION. IF IRTRN
C                IS SET <0, ODEX2 WILL RETURN TO THE CALLING PROGRAM.
C
C      ----- CONTINUOUS OUTPUT (IF IOUT=2): -----
C                DURING CALLS TO "SOLOUT", A CONTINUOUS SOLUTION
C                FOR THE INTERVAL [XOLD,X] IS AVAILABLE THROUGH
C                THE REAL*8 FUNCTION
C                >>> CONTX2(I,S) <<<
C                WHICH PROVIDES AN APPROXIMATION TO THE I-TH
C                COMPONENT OF THE SOLUTION AT THE POINT S. THE VALUE
C                S SHOULD LIE IN THE INTERVAL [XOLD,X].
C
C      .....
C
C      WORK      ARRAY OF WORKING SPACE OF LENGTH "LWORK".
C                SERVES AS WORKING SPACE FOR ALL VECTORS.
C                "LWORK" MUST BE AT LEAST
C                N*(2*KM+6)+5*KM+10+KM*(2*KM+3)*NRDENS
C                WHERE NRDENS=IWORK(8) (SEE BELOW) AND
C                KM=9 IF IWORK(2)=0
C                KM=IWORK(2) IF IWORK(2).GT.0
C                WORK(1),...,WORK(10) SERVE AS PARAMETERS
C                FOR THE CODE. FOR STANDARD USE, SET THESE
C                PARAMETERS TO ZERO BEFORE CALLING.
C
C      .....
C
C      IWORK     INTEGER WORKING SPACE OF LENGTH "LIWORK".
C                "LIWORK" MUST BE AT LEAST
C                KM+9+NRDENS
C                IWORK(1),...,IWORK(9) SERVE AS PARAMETERS
C                FOR THE CODE. FOR STANDARD USE, SET THESE
C                PARAMETERS TO ZERO BEFORE CALLING.
C
C      .....
C
C      LRCONT    DECLARED LENGTH OF COMMON BLOCK
C                >>> COMMON /CONTR2/RCONT(LRCONT) <<<

```

```

C          WHICH MUST BE DECLARED IN THE CALLING PROGRAM.
C          "LRCONT" MUST BE AT LEAST
C              ( 2 * KM + 6 ) * NRDENS + 2
C          WHERE KM=IWORK(2) AND NRDENS=IWORK(8) (SEE BELOW).
C
C          LICONT    DECLARED LENGTH OF COMMON BLOCK
C              >>> COMMON /CONTI2/ICONT(LICONT) <<<
C          WHICH MUST BE DECLARED IN THE CALLING PROGRAM.
C          "LICONT" MUST BE AT LEAST
C              NRDENS + 2
C          THESE COMMON BLOCKS ARE USED FOR STORING THE COEFFICIENTS
C          OF THE CONTINUOUS SOLUTION AND MAKES THE CALLING LIST FOR
C          THE FUNCTION "CONTX2" AS SIMPLE AS POSSIBLE.
C
C          .....
C
C          WORK(3)    STEP SIZE IS REDUCED BY FACTOR WORK(3), IF DURING THE
C                    COMPUTATION OF THE EXTRAPOLATION TABLEAU DIVERGENCE
C                    IS OBSERVED; DEFAULT 0.5.
C
C          .....
C
C          IWORK(3)   SWITCH FOR THE STEP SIZE SEQUENCE (EVEN NUMBERS ONLY)
C                    IF IWORK(3).EQ.1 THEN 2,4,6,8,10,12,14,16,...
C                    IF IWORK(3).EQ.2 THEN 2,4,8,12,16,20,24,28,...
C                    IF IWORK(3).EQ.3 THEN 2,4,6,8,12,16,24,32,...
C                    IF IWORK(3).EQ.4 THEN 2,6,10,14,18,22,26,30,...
C                    THE DEFAULT VALUE IS IWORK(3)=1 IF IOUT.LE.1;
C                    THE DEFAULT VALUE IS IWORK(3)=4 IF IOUT.GE.2.
C
C          .....
C
C          IWORK(7)   DETERMINES THE DEGREE OF INTERPOLATION FORMULA
C                    MU = 2 * KAPPA - IWORK(7) + 1
C                    IWORK(7) SHOULD LIE BETWEEN 1 AND 8
C                    DEFAULT IWORK(7)=6 (IF IWORK(7)=0).
C
C          .....

```

Driver for the Code RETARD

We consider the delay equation (II.17.14) with initial values and initial functions given there. This is a 3-dimensional problem, but only the second component is used with retarded argument (hence $\text{NRDENS}=1$). We require that the points 1, 2, 3, ..., 9, 10, 20 (points of discontinuity of the derivatives of the solution) are hitten exactly by the integration routine.

```

      IMPLICIT REAL*8 (A-H,O-Z)
      PARAMETER (NDGL=3,NGRID=11,LWORK=8*NDGL+11+NGRID,LIWORK=10)
      PARAMETER (NRDENS=1,LRCONT=500,LICONT=NRDENS+1)
      DIMENSION Y(NDGL),WORK(LWORK),IWORK(LIWORK)
      COMMON/STATRE/NFCN,NSTEP,NACCP,NREJCT
      COMMON /CORER/RCONT(LRCONT)
      COMMON /COREI/ICONT(LICONT)
      EXTERNAL FCN,SOLOUT
C --- DIMENSION OF THE SYSTEM
      N=NDGL
C --- OUTPUT ROUTINE IS USED DURING INTEGRATION

```



```

      IOUT=1
C --- INITIAL VALUES AND ENDPOINT OF INTEGRATION
      X=0.0D0
      Y(1)=5.0D0
      Y(2)=0.1D0
      Y(3)=1.0D0
      XEND=40.D0
C --- REQUIRED (RELATIVE AND ABSOLUTE) TOLERANCE
      ITOL=0
      RTOL=1.0D-5
      ATOL=RTOL
C --- DEFAULT VALUES FOR PARAMETERS
      DO 10 I=1,10
        IWORK(I)=0
      10   WORK(I)=0.D0
C --- SECOND COMPONENT USES RETARDED ARGUMENT
      IWORK(5)=NRDENS
      ICONT(2)=2
C --- USE AS GRID-POINTS
      IWORK(6)=NGRID
      DO 12 I=1,NGRID-1
      12   WORK(10+I)=I
      WORK(10+NGRID)=20.D0
C --- CALL OF THE SUBROUTINE RETARD
      CALL RETARD(N,FCN,X,Y,XEND,
+              RTOL,ATOL,ITOL,
+              SOLOUT,IOUT,
+              WORK,LWORK,IWORK,LIWORK,LRCONT,LICONT,IDID)
C --- PRINT FINAL SOLUTION
      WRITE (6,99) Y(1),Y(2),Y(3)
      99   FORMAT(1X,'X = XEND      Y =',3E18.10)
C --- PRINT STATISTICS
      WRITE (6,91) RTOL,NFCN,NSTEP,NACCPY,NREJCT
      91   FORMAT('      tol=',D8.2,'      fcn=',I5,'      step=',I4,
+              '      accpt=',I4,'      reject=',I3)
      STOP
      END

C
C
      SUBROUTINE SOLOUT (NR,XOLD,X,Y,N,IRTRN)
C --- PRINTS SOLUTION AT EQUIDISTANT OUTPUT-POINTS
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION Y(N)
      EXTERNAL PHI
      COMMON /INTERN/XOUT
      IF (NR.EQ.1) THEN
        WRITE (6,99) X,Y(1),NR-1
        XOUT=X+5.D0
      ELSE
      10   CONTINUE
        IF (X.GE.XOUT) THEN
          WRITE (6,99) X,Y(1),NR-1
          XOUT=XOUT+5.D0
          GOTO 10
        END IF
      END IF
      99   FORMAT(1X,'X =',F6.2,'      Y =',E18.10,'      NSTEP =',I4)
      RETURN
      END
C

```

```

      SUBROUTINE FCN(N,X,Y,F)
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION Y(N),F(N)
      EXTERNAL PHI
      Y2L1=YLAG(2,X-1.DO,PHI)
      Y2L10=YLAG(2,X-10.DO,PHI)
      F(1)=-Y(1)*Y2L1+Y2L10
      F(2)=Y(1)*Y2L1-Y(2)
      F(3)=Y(2)-Y2L10
      RETURN
      END
C
      FUNCTION PHI(I,X)
      IMPLICIT REAL*8 (A-H,O-Z)
      IF (I.EQ.2) PHI=0.1D0
      RETURN
      END

```

The result, obtained on an Apollo workstation, is the following:

```

X = 0.00      Y = 0.5000000000E+01      NSTEP = 0
X = 5.00      Y = 0.2533855892E+00      NSTEP = 18
X = 10.00     Y = 0.3328560326E+00      NSTEP = 32
X = 15.29     Y = 0.4539376456E+01      NSTEP = 40
X = 20.00     Y = 0.1706635702E+00      NSTEP = 52
X = 25.22     Y = 0.2524799457E+00      NSTEP = 62
X = 30.48     Y = 0.5134266860E+01      NSTEP = 68
X = 35.10     Y = 0.3610797907E+00      NSTEP = 78
X = 40.00     Y = 0.9125544555E-01      NSTEP = 89
X = XEND      Y = 0.9125544555E-01      0.2029882456E-01  0.5988445730E+01
      tol=0.10E-04      fcn= 586 step= 97 accpt= 89 reject= 8

```

Subroutine RETARD

Modification of the code DOPRI5 for delay differential equations (see Section II.17). The missing parts of the description are identical to those of DOPRI5.

```

      SUBROUTINE RETARD(N,FCN,X,Y,XEND,
+                      RTOL,ATOL,ITOL,
+                      SOLOUT,IOUT,
+                      WORK,LWORK,IWORK,LIWORK,LRCONT,LICONT,IDID)
C -----
C   NUMERICAL SOLUTION OF A SYSTEM OF FIRST ORDER DELAY
C   ORDINARY DIFFERENTIAL EQUATIONS  Y'(X)=F(X,Y(X),Y(X-A),...).
C   THIS CODE IS BASED ON AN EXPLICIT RUNGE-KUTTA METHOD OF
C   ORDER (4)5 DUE TO DORMAND & PRINCE (WITH STEPSIZE CONTROL
C   AND DENSE OUTPUT).
C   .....
C
C   VERSION OF APRIL 24, 1992
C   .....
C
C   FCN          NAME (EXTERNAL) OF SUBROUTINE COMPUTING THE RIGHT-
C                HAND-SIDE OF THE DELAY EQUATION, E.G.,

```

```

C          SUBROUTINE FCN(N,X,Y,F)
C          REAL*8 X,Y(N),F(N)
C          EXTERNAL PHI
C          F(1)=(1.4D0-YLAG(1,X-1.D0,PHI))*Y(1)
C          F(2)=...      ETC.
C          FOR AN EXPLICATION OF YLAG SEE BELOW.
C          DO NOT USE YLAG(I,X-0.D0,PHI) !
C          THE INITIAL FUNCTION HAS TO BE SUPPLIED BY:
C          FUNCTION PHI(I,X)
C          REAL*8 PHI,X
C          WHERE I IS THE COMPONENT AND X THE ARGUMENT
C          .....
C          Y(N)      INITIAL VALUES FOR Y (MAY BE DIFFERENT FROM PHI (I,X),
C                   IN THIS CASE IT IS HIGHLY RECOMMENDED TO SET IWORK(6)
C                   AND WORK(11),..., SEE BELOW)
C          .....
C          ----- CONTINUOUS OUTPUT: -----
C                   DURING CALLS TO "SOLOUT" AS WELL AS TO "FCN", A
C                   CONTINUOUS SOLUTION IS AVAILABLE THROUGH THE FUNCTION
C                   >>> YLAG(I,S,PHI) <<<
C                   WHICH PROVIDES AN APPROXIMATION TO THE I-TH
C                   COMPONENT OF THE SOLUTION AT THE POINT S. THE VALUE S
C                   HAS TO LIE IN AN INTERVAL WHERE THE NUMERICAL SOLUTION
C                   IS ALREADY COMPUTED. IT DEPENDS ON THE SIZE OF LRCONT
C                   (SEE BELOW) HOW FAR BACK THE SOLUTION IS AVAILABLE.
C          IOUT      SWITCH FOR CALLING THE SUBROUTINE SOLOUT:
C                   IOUT=0: SUBROUTINE IS NEVER CALLED
C                   IOUT=1: SUBROUTINE IS USED FOR OUTPUT.
C          WORK      ARRAY OF WORKING SPACE OF LENGTH "LWORK".
C                   "LWORK" MUST BE AT LEAST 8*N+11+NGRID
C                   WHERE NGRID=IWORK(6)
C          .....
C          LRCONT    DECLARED LENGTH OF COMMON BLOCK
C                   >>> COMMON /CORER/RCONT(LRCONT) <<<
C                   WHICH MUST BE DECLARED IN THE CALLING PROGRAM.
C                   "LRCONT" MUST BE SUFFICIENTLY LARGE. IF THE DENSE
C                   OUTPUT OF MXST BACK STEPS HAS TO BE STORED, IT MUST
C                   BE AT LEAST
C                   MXST * ( 5 * NRDENS + 2 )
C                   WHERE NRDENS=IWORK(5) (SEE BELOW).
C          LICONT    DECLARED LENGTH OF COMMON BLOCK
C                   >>> COMMON /COREI/ICONT(LICONT) <<<
C                   WHICH MUST BE DECLARED IN THE CALLING PROGRAM.
C                   "LICONT" MUST BE AT LEAST
C                   NRDENS + 1
C                   THESE COMMON BLOCKS ARE USED FOR STORING THE COEFFICIENTS
C                   OF THE CONTINUOUS SOLUTION AND MAKES THE CALLING LIST FOR
C                   THE FUNCTION "CONTD5" AS SIMPLE AS POSSIBLE.
C          .....
C          WORK(11),...,WORK(10+NGRID) PRESCRIBED POINTS, WHICH THE
C          INTEGRATION METHOD HAS TO TAKE AS GRID-POINTS
C          X < WORK(11) < WORK(12) < ... < WORK(10+NGRID) <= XEND
C          .....

```

```

C
C   IWORK(5) = NRDENS = NUMBER OF COMPONENTS, FOR WHICH DENSE OUTPUT
C               IS REQUIRED (EITHER BY "SOLOUT" OR BY "FCN");
C               DEFAULT VALUE (FOR IWORK(5)=0) IS IWORK(5)=N;
C               FOR 0 < NRDENS < N THE COMPONENTS (FOR WHICH DENSE
C               OUTPUT IS REQUIRED) HAVE TO BE SPECIFIED IN
C               ICONT(2),...,ICONT(NRDENS+1);
C               FOR NRDENS=N THIS IS DONE BY THE CODE.
C
C   IWORK(6) = NGRID = NUMBER OF PRESCRIBED POINTS IN THE
C               INTEGRATION INTERVAL WHICH HAVE TO BE GRID-POINTS
C               IN THE INTEGRATION. USUALLY, AT THESE POINTS THE
C               SOLUTION OR ONE OF ITS DERIVATIVE HAS A DISCONTINUITY.
C               DEFINE THESE POINTS IN WORK(11),...,WORK(10+NGRID)
C               DEFAULT VALUE: IWORK(6)=0
C   .....
C
C   IDID      REPORTS ON SUCCESSFULNESS UPON RETURN:
C               IDID= 1 COMPUTATION SUCCESSFUL,
C               IDID= 2 COMPUT. SUCCESSFUL (INTERRUPTED BY SOLOUT)
C               IDID=-1 INPUT IS NOT CONSISTENT,
C               IDID=-2 LARGER NMAX IS NEEDED,
C               IDID=-3 STEP SIZE BECOMES TOO SMALL.
C               IDID=-4 PROBLEM IS PROBABLY STIFF (INTERRUPTED).
C               IDID=-5 COMPUT. INTERRUPTED BY YLAG
C   .....

```

Bibliography

This bibliography includes the publications referred to in the text. Italic numbers in square brackets following a reference indicate the sections where the reference is cited.

N.H. Abel (1826): *Untersuchungen über die Reihe:*

$$1 + \frac{m}{1}x + \frac{m(m-1)}{1 \cdot 2}x^2 + \frac{m(m-1)(m-2)}{1 \cdot 2 \cdot 3}x^3 + \dots \text{ u.s.w.}$$

Crelle J. f. d. r. u. angew. Math. (in zwanglosen Heften), Vol.1, p.311-339. [III.8]

N.H. Abel (1827): *Ueber einige bestimmte Integrale*. Crelle J. f. d. r. u. angew. Math., Vol.2, p.22-30. [I.11]

L. Abia & J.M. Sanz-Serna (1993): *Partitioned Runge-Kutta methods for separable Hamiltonian problems*. Math. Comp., Vol.60, p.617-634. [II.16]

L. Abia, see also J.M. Sanz-Serna & L. Abia.

M. Abramowitz & I.A. Stegun (1964): *Handbook of mathematical functions*. Dover, 1000 pages. [II.7], [II.8], [II.9]

J.C. Adams (1883): see F.Bashforth (1883).

R.C. Aiken ed. (1985): *Stiff computation*. Oxford, Univ. Press, 462pp. [Appendix]

A.C. Aitken (1932): *On interpolation by iteration of proportional parts, without the use of differences*. Proc. Edinburgh Math. Soc. Second ser., Vol.3, p.56-76. [II.9]

J. Albrecht (1955): *Beiträge zum Runge-Kutta-Verfahren*. ZAMM, Vol.35, p.100-110. [II.13], [II.14]

P. Albrecht (1978): *Explicit, optimal stability functionals and their application to cyclic discretization methods*. Computing, Vol.19, p.233-249. [III.8]

P. Albrecht (1979): *Die numerische Behandlung gewöhnlicher Differentialgleichungen*. Akademie Verlag, Berlin; Hanser Verlag, München. [III.8]

P. Albrecht (1985): *Numerical treatment of O.D.E.s.: The theory of A-methods*. Numer. Math., Vol.47, p.59-87. [III.8]

V.M. Alekseev (1961): *An estimate for the perturbations of the solution of ordinary differential equations (Russian)*. Vestn. Mosk. Univ., Ser.I, Math. Meh, 2, p.28-36. [I.14]

J.le Rond d'Alembert (1743): *Traité de dynamique, dans lequel les loix de l'équilibre & du mouvement des corps sont réduites au plus petit nombre possible, & démontrées d'une manière nouvelle, & où l'on donne un principe général pour trouver le mouvement de*

- plusieurs corps qui agissent les uns sur les autres, d'une manière quelconque.* à Paris, MDCCXLIII, 186p., 70 figs. [I.6]
- J.le Rond d'Alembert (1747): *Recherches sur la courbe que forme une corde tenduë mise en vibration.* Hist. de l'Acad. Royale de Berlin, Tom.3, Année MDCCXLVII, publ. 1749, p.214-219 et 220-249. [I.6]
- J.le Rond d'Alembert (1748): *Suite des recherches sur le calcul intégral, quatrième partie: Méthodes pour intégrer quelques équations différentielles.* Hist. Acad. Berlin, Tom.IV, p.275-291. [I.4]
- R.F. Arenstorf (1963): *Periodic solutions of the restricted three body problem representing analytic continuations of Keplerian elliptic motions.* Amer. J. Math., Vol.LXXXV, p.27-35. [II.0]
- V.I. Arnol'd (1974): *Mathematical methods of classical mechanics.* Nauka, Moscow; French transl. Mir 1976; Engl. transl. Springer-Verlag 1978 (2nd edition 1989). [I.14]
- C. Arzelà (1895): *Sulle funzioni di linee.* Memorie dell. R. Accad. delle Sc. di Bologna, 5e serie, Vol.V, p.225-244, see also: Vol.V, p.257-270, Vol.VI, (1896), p.131-140. [I.7]
- U.M. Ascher, R.M.M. Mattheij & R.D. Russel (1988): *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations.* Prentice Hall, Englewood Cliffs. [I.15]
- L.S. Baca, see L.F. Shampine & L.S. Baca, L.F. Shampine, L.S. Baca & H.-J. Bauer.
- H.F. Baker (1905): *Alternants and continuous groups.* Proc. London Math. Soc., Second Ser., Vol.3, p.24-47. [II.16]
- N. Bakhvalov (1976): *Méthodes numériques.* Editions Mir, Moscou 600pp., russian edition 1973. [I.9]
- F. Bashforth (1883): *An attempt to test the theories of capillary action by comparing the theoretical and measured forms of drops of fluid. With an explanation of the method of integration employed in constructing the tables which give the theoretical form of such drops,* by J.C.Adams. Cambridge Univ. Press. [III.1]
- R.H. Battin (1976): *Resolution of Runge-Kutta-Nyström condition equations through eighth order.* AIAA J., Vol.14, p.1012-1021. [II.14]
- F.L. Bauer, H. Rutishauser & E. Stiefel (1963): *New aspects in numerical quadrature.* Proc. of Symposia in Appl. Math., Vol.15, p.199-218, Am. Math. Soc. [II.9]
- H.-J. Bauer, see L.F. Shampine, L.S. Baca & H.-J. Bauer.
- P.A. Beentjes & W.J. Gerritsen (1976): *Higher order Runge-Kutta methods for the numerical solution of second order differential equations without first derivatives.* Report NW 34/76, Math. Centrum, Amsterdam. [II.14]
- H. Behnke & F. Sommer (1962): *Theorie der analytischen Funktionen einer komplexen Veränderlichen.* Zweite Auflage. Springer Verlag, Berlin-Göttingen-Heidelberg. [III.2]
- A. Bellen (1984): *One-step collocation for delay differential equations.* J. Comput. Appl. Math., Vol.10, p.275-283. [II.17]
- A. Bellen & M. Zennaro (1985): *Numerical solution of delay differential equations by uniform corrections to an implicit Runge-Kutta method.* Numer. Math., Vol.47, p.301-316. [II.17]

- R. Bellman & K.L. Cooke (1963): *Differential-Difference equations*. Academic Press, 462pp. [II.17]
- I. Bendixson (1893): *Sur le calcul des intégrales d'un système d'équations différentielles par des approximations successives*. Stock. Akad. Öfversigt Förh., Vol.50, p.599-612. [I.8]
- I. Bendixson (1901): *Sur les courbes définies par des équations différentielles*. Acta Mathematica, Vol.24, p.1-88. [I.16]
- I.S. Berezin & N.P Zhidkov (1965): *Computing methods (Metody vychislenii)*. 2 Volumes, Fizmatgiz Moscow, Engl. transl.: Pergamon Press, 464 & 679pp. [I.1]
- Dan. Bernoulli (1728): *Observationes de seriebus quae formantur ex additione vel subtractione quacunque terminorum se mutuo consequentium, ubi praesertim earundem insignis usus pro inveniendis radicibus omnium aequationum algebraicarum ostenditur*. Comm. Acad. Sci. Imperialis Petrop., Tom.III, 1728 (1732), p.85-100; Werke, Bd.2, p.49-70. [III.3]
- Dan. Bernoulli (1732): *Theoremata de oscillationibus corporum filo flexili connexorum et catenae verticaliter suspensae*. Comm. Acad. Sci. Imperialis Petrop., Tom.VI, ad annum MDCCXXXII & MDCCXXXIII, p.108-122. [I.6]
- Dan. Bernoulli (1760): *Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir*. Hist. et Mém. de l'Acad. Roy. Sciences Paris, 1760, p.1-45; Werke Bd. 2, p.235-267. [II.17]
- Jac. Bernoulli (1690): *Analysis problematis ante hac propositi, de inventione lineae descensus a corpore gravi percurrendae uniformiter, sic ut temporibus aequalibus aequales altitudines emetiatur: & alterius cujusdam Problematis Propositio*. Acta Erudit. Lipsiae, Anno MDCLXXX, p. 217-219. [I.3]
- Jac. Bernoulli (1695): *Explicationes, Annotationes & Additiones ad ea, quae in Actis sup. anni de Curva Elastica, Isochrone Paracentrica, & Velaria, hinc inde memorata, & partim controversa leguntur; ubi de Linea mediarum directionum, aliisque novis*. Acta Erudit. Lipsiae, Anno MDCXCV, p. 537-553. [I.3]
- Jac. Bernoulli (1697): *Solutio Problematum Fraternalium, Peculiari Programme Cal. Jan. 1697 Groningae, nec non Actorum Lips. mense Jun. & Dec. 1696, & Febr. 1697 propositorum: una cum Propositione reciproca aliorum*. Acta Erud. Lips. MDCXCVII, p.211-217. [I.2]
- Joh. Bernoulli (1691): *Solutio problematis funicularii, exhibita à Johanne Bernoulli, Basil. Med. Cand.*. Acta Erud. Lips. MDCXCI, p.274, Opera Omnia, Vol.I, p.48-51, Lausannae & Genevae 1742. [I.3]
- Joh. Bernoulli (1696): *Problema novum Mathematicis propositorum*. Acta Erud. Lips. MDCXCVI, p.269, Opera Omnia, Vol.I, p.161 and 165, Lausannae & Genevae 1742. [I.2]
- Joh. Bernoulli (1697): *De Conoidibus et Sphaeroidibus quaedam. Solutio analytica Aequationis in Actis A. 1695, pag. 553 propositae*. Acta Erud. Lips., MDCXCVII, p.113-118. Opera Omnia, Vol.I, p.174-179. [I.3]
- Joh. Bernoulli (1697b): *Solutioque Problematis a se in Actis 1696, p.269, propositi, de invenienda Linea Brachystochrona*. Acta Erud.Lips. MDCXCVII, p.206, Opera Omnia, Vol.I, p.187-193. [I.2]
- Joh. Bernoulli (1727): *Meditationes de chordis vibrantibus . . .* Comm. Acad. Sci. Imperialis Petrop., Tom.III, p.13; Opera, Vol.III, p.198-210. [I.6]

- J. Berntsen & T.O. Espelid (1991): *Error estimation in automatic quadrature routines*. ACM Trans. on Math. Software, Vol.17, p.233-255. [II.10]
- D.G. Bettis (1973): *A Runge-Kutta Nyström algorithm*. Celestial Mechanics, Vol.8, p.229-233. [II.14]
- L. Bieberbach (1923): *Theorie der Differentialgleichungen*. Grundlehren Bd.VI, Springer Verlag. [II.3]
- L. Bieberbach (1951): *On the remainder of the Runge-Kutta formula in the theory of ordinary differential equations*. ZAMP, Vol.2, p.233-248. [II.3]
- J. Binney (1981): *Resonant excitation of motion perpendicular to galactic planes*. Mon. Not. R. astr. Soc., Vol.196, p.455-467. [II.16]
- J. Binney & S. Tremaine (1987): *Galactic dynamics*. Princeton Univ. Press, 733pp. [II.16]
- J.B. Biot (1804): *Mémoire sur la propagation de la chaleur, lu à la classe des sciences math. et phys. de l'Institut national*. Bibl. britann. Sept 1804, 27, p.310. [I.6]
- G. Birkhoff & R.S. Varga (1965): *Discretization errors for well-set Cauchy problems I*. Journal of Math. and Physics, Vol.XLIV, p.1-23. [I.13]
- H.G. Bock (1981): *Numerical treatment of inverse problems in chemical reaction kinetics*. In: Modelling of chemical reaction systems, ed. by K.H. Ebert, P. Deuflhard & W. Jäger, Springer Series in Chem. Phys., Vol.18, p.102-125. [II.6]
- H.G. Bock & J. Schlöder (1981): *Numerical solution of retarded differential equations with statedependent time lages*. ZAMM, Vol.61, p.269-271. [II.17]
- P. Bogacki & L.F. Shampine (1989): *An efficient Runge-Kutta (4,5) pair*. SMU Math Rept 89-20. [II.6]
- N. Bogoliuboff, see N. Kryloff & N. Bogoliuboff.
- R.W. Brankin, I. Gladwell, J.R. Dormand, P.J. Prince & W.L. Seward (1989): *Algorithm 670. A Runge-Kutta-Nyström code*. ACM Trans. Math. Softw., Vol.15, p.31-40. [II.14]
- R.W. Brankin, see also I. Gladwell, L.F. Shampine & R.W. Brankin.
- P.N. Brown, G.D. Byrne & A.C. Hindmarsh (1989): *VODE: a variable-coefficient ODE solver*. SIAM J. Sci. Stat. Comput., Vol.10, p.1038-1051. [III.7]
- H. Brunner & P.J. van der Houwen (1986): *The numerical solution of Volterra equations*. North-Holland, Amsterdam, 588pp. [II.17]
- R. Bulirsch & J. Stoer (1964): *Fehlerabschätzungen und Extrapolation mit rationalen Funktionen bei Verfahren vom Richardson-Typus*. Num. Math., Vol.6, p.413-427. [II.9]
- R. Bulirsch & J. Stoer (1966): *Numerical treatment of ordinary differential equations by extrapolation methods*. Num. Math., Vol.8, p.1-13. [II.9]
- K. Burrage (1985): *Order and stability properties of explicit multivalued methods*. Appl. Numer. Anal., Vol.1, p.363-379. [III.8]
- K. Burrage & J.C. Butcher (1980): *Non-linear stability of a general class of differential equation methods*. BIT, Vol.20, p.185-203. [III.8]
- K. Burrage & P. Moss (1980): *Simplifying assumptions for the order of partitioned multivalued methods*. BIT, Vol.20, p.452-465. [III.8]

- J.C. Butcher (1963): *Coefficients for the study of Runge-Kutta integration processes*. J. Austral. Math. Soc., Vol.3, p.185-201. [II.2]
- J.C. Butcher (1963a): *On the integration process of A. Huřa*. J. Austral. Math. Soc., Vol.3, p.202-206. [II.2]
- J.C. Butcher (1964a): *Implicit Runge-Kutta Processes*. Math. Comput., Vol.18, p.50-64. [II.7], [II.16]
- J.C. Butcher (1964b): *On Runge-Kutta processes of high order*. J. Austral. Math. Soc., Vol. IV, Part2, p.179-194. [II.1], [II.5]
- J.C. Butcher (1964c): *Integration processes based on Radau quadrature formulas*. Math. Comput., Vol.18, p.233-244. [II.7]
- J.C. Butcher (1965a): *A modified multistep method for the numerical integration of ordinary differential equations*. J. ACM, Vol.12, p.124-135. [III.8]
- J.C. Butcher (1965b): *On the attainable order of Runge-Kutta methods*. Math. of Comp., Vol.19, p.408-417. [II.5]
- J.C. Butcher (1966): *On the convergence of numerical solutions to ordinary differential equations*. Math. Comput., Vol.20, p.1-10. [III.4], [III.8]
- J.C. Butcher (1969): *The effective order of Runge-Kutta methods*. in: Conference on the numerical solution of differential equations, Lecture Notes in Math., Vol.109, p.133-139. [II.12]
- J.C. Butcher (1981): *A generalization of singly-implicit methods*. BIT, Vol.21, p.175-189. [III.8]
- J.C. Butcher (1984): *An application of the Runge-Kutta space*. BIT, Vol.24, p.425-440. [II.12], [III.8]
- J.C. Butcher (1985a): *General linear method: a survey*. Appl. Num. Math., Vol.1, p.273-284. [III.8]
- J.C. Butcher (1985b): *The non-existence of ten stage eighth order explicit Runge-Kutta methods*. BIT, Vol.25, p.521-540. [II.5]
- J.C. Butcher (1987): *The numerical analysis of ordinary differential equations. Runge-Kutta and general linear methods*. John Wiley & Sons, Chichester, 512pp. [II.16]
- J.C. Butcher, see also K. Burrage & J.C. Butcher.
- G.D. Byrne & A.C. Hindmarsh (1975): *A polyalgorithm for the numerical solution of ordinary differential equations*. ACM Trans. on Math. Software, Vol.1, No.1, p.71-96. [III.6], [III.7]
- G.D. Byrne & R.J. Lambert (1966): *Pseudo-Runge-Kutta methods involving two points*. J. Assoc. Comput. Mach., Vol.13, p.114-123. [III.8]
- G.D. Byrne, see also P.N. Brown, G.D. Byrne & A.C. Hindmarsh.
- R. Caira, C. Costabile & F. Costabile (1990): *A class of pseudo Runge-Kutta methods*. BIT, Vol.30, p.642-649. [III.8]
- M. Calvé & R. Vaillancourt (1990): *Interpolants for Runge-Kutta pairs of order four and five*. Computing, Vol.45, p.383-388. [II.6]

- M. Calvo, J.I. Montijano & L. Rández (1990): *A new embedded pair of Runge-Kutta formulas of orders 5 and 6*. Computers Math. Applic., Vol.20, p.15-24. [II.6]
- M. Calvo, J.I. Montijano & L. Rández (1992): *New continuous extensions for the Dormand and Prince RK method*. In: Computational ordinary differential equations, ed. by J.R. Cash & I. Gladwell, Clarendon Press, Oxford, p.135-164. [II.6]
- M.P. Calvo & J.M. Sanz-Serna (1992): *Order conditions for canonical Runge-Kutta-Nyström methods*. BIT, Vol.32, p.131-142. [II.16]
- M.P. Calvo & J.M. Sanz-Serna (1992b): *High order symplectic Runge-Kutta-Nyström methods*. SIAM J. Sci. Stat. Comput., Vol.14 (1993), p.1237-1252. [II.16]
- M.P. Calvo & J.M. Sanz-Serna (1992c): *Reasons for a failure. The integration of the two-body problem with a symplectic Runge-Kutta method with step changing facilities*. Intern. Conf. on Differential Equations, Vol. 1, 2 (Barcelona, 1991), 93-102, World Sci. Publ., River Edge, NJ, 1993. [II.16]
- J.M. Carnicer (1991): *A lower bound for the number of stages of an explicit continuous Runge-Kutta method to obtain convergence of given order*. BIT, Vol.31, p.364-368. [II.6]
- E. Cartan (1899): *Sur certaines expressions différentielles et le problème de Pfaff*. Ann. Ecol. Normale, Vol.16, p.239-332, Oeuvres partie II, p.303-396. [I.14]
- A.L. Cauchy (1824): *Résumé des Leçons données à l'Ecole Royale Polytechnique. Suite du Calcul Infinitésimal*; published: Equations différentielles ordinaires, ed. Chr. Gilain, Johnson 1981. [I.2], [I.7], [I.9], [II.3], [II.7]
- A.L. Cauchy (1831): *Sur la mecanique celeste et sur un nouveau calcul appelé calcul des limites*. lu à l'acad. de Turin le 11 oct 1831; also: exerc. d'anal. et de physique math, 2, Paris 1841; oeuvres (2), 12. [III.3]
- A.L. Cauchy (1839-42): *Several articles in Comptes Rendus de l'Acad. des Sciences de Paris*. (Aug. 5, Nov. 21, 1839, June 29, Oct. 26, 1840, etc). [I.8]
- A. Cayley (1857): *On the theory of the analytic forms called trees*. Phil. Magazine, Vol.XIII, p.172-176, Mathematical Papers, Vol.3, Nr.203, p.242-246. [II.2]
- A. Cayley (1858): *A memoir on the theory of matrices*. Phil. Trans. of Royal Soc. of London, Vol.CXLVIII, p.17-37, Mathematical Papers, Vol.2, Nr.152, p.475.
- F. Ceschino (1961): *Modification de la longueur du pas dans l'intégration numérique par les méthodes à pas liés*. Chiffres, Vol.2, p.101-106. [II.4], [III.5]
- F. Ceschino (1962): *Evaluation de l'erreur par pas dans les problèmes différentiels*. Chiffres, Vol.5, p.223-229. [II.4]
- F. Ceschino & J. Kuntzmann (1963): *Problèmes différentiels de conditions initiales (méthodes numériques)*. Dunod Paris, 372pp.; english translation: Numerical solutions of initial value problems, Prentice Hall 1966 [II.5], [II.7]
- P.J. Channell & C. Scovel (1990): *Symplectic integration of Hamiltonian systems*. Nonlinearity, Vol.3, p.231-259. [II.16]
- A.C. Clairaut (1734): *Solution de plusieurs problèmes où il s'agit de trouver des courbes dont la propriété consiste dans une certaine relation entre leurs branches, exprimée par une Equation donnée*. Mémoires de Math. et de Phys. de l'Acad. Royale des Sciences, Paris, Année MDCCXXXIV, p.196-215. [I.2]

- L. Collatz (1951): *Numerische Behandlung von Differentialgleichungen*. Grundlehren Band LX. Springer Verlag, 458pp; second edition 1955; third edition and english translation 1960. [II.7]
- L. Collatz (1967): *Differentialgleichungen. Eine Einführung unter besonderer Berücksichtigung der Anwendungen*. Leitfäden der angewandten Mathematik, Teubner 226pp. English translation: *Differential equations. An introduction with applications*, Wiley, 372pp., (1986). [I.15]
- P. Collet & J.P. Eckmann (1980): *Iterated maps on the interval as dynamical systems*. Birkhäuser, 248pp. [I.16]
- K.L. Cooke, see R. Bellman & K.L. Cooke.
- G.J. Cooper (1978): *The order of convergence of general linear methods for ordinary differential equations*. SIAM, J. Numer. Anal., Vol.15, p.643-661. [III.8]
- G.J. Cooper (1987): *Stability of Runge-Kutta methods for trajectory problems*. IMA J. Numer. Anal., Vol.7, p.1-13. [II.16]
- G.J. Cooper & J.H. Verner (1972): *Some explicit Runge-Kutta methods of high order*. SIAM J. Numer. Anal., Vol.9, p.389-405. [II.5]
- S.A. Corey (1906): *A method of approximation*. Amer. Math. Monthly, Vol.13, p.137-140. [II.9]
- C. Costabile, see R. Caira, C. Costabile & F. Costabile.
- F. Costabile, see R. Caira, C. Costabile & F. Costabile.
- C.A. de Coulomb (1785): *Théorie des machines simples, en ayant égard au frottement de leurs parties, et a la roideur des cordages*. Pièce qui a remporté le Prix double de l'Académie des Sciences pour l'année 1781. Mémoires des Savans Etrangers, tome X, p. 163-332; réimprimé 1809 chez Bachelier, Paris. [II.6]
- P.H. Cowell & A.C.D. Crommelin (1910): *Investigation of the motion of Halley's comet from 1759 to 1910*. Appendix to Greenwich Observations for 1909, Edinburgh, p.1-84. [III.10]
- J.W. Craggs, see A.R. Mitchell & J.W. Craggs.
- D.M. Creedon & J.J.H. Miller (1975): *The stability properties of q-step backward-difference schemes*. BIT, Vol.15, p.244-249. [III.3]
- A.C.D. Crommelin, see P.H. Cowell & A.C.D. Crommelin.
- M. Crouzeix (1975): *Sur l'approximation des équations différentielles opérationnelles linéaires par des méthodes de Runge-Kutta*. Thèse d'état, Univ. Paris 6, 192pp. [II.2], [II.7]
- M. Crouzeix & F.J. Lisbona (1984): *The convergence of variable-stepsize, variable formula, multistep methods*. SIAM J. Num. Anal., Vol.21, p.512-534. [III.5]
- C.W. Cryer (1971): *A proof of the instability of backward-difference multistep methods for the numerical integration of ordinary differential equations*. Tech. Rep. No.117, Comp. Sci. Dept., Univ. of Wisconsin, p.1-52. [III.3]
- C.W. Cryer (1972): *On the instability of high order backward-difference multistep methods*. BIT, Vol.12, p.17-25. [III.3]
- W.J. Cunningham (1954): *A nonlinear differential-difference equation of growth*. Proc. Mat. Acad. Sci., USA, Vol.40, p.708-713. [II.17]

- A.R. Curtis (1970): *An eighth order Runge-Kutta process with eleven function evaluations per step*. Numer. Math., Vol.16, p.268-277. [II.5]
- A.R. Curtis (1975): *High-order explicit Runge-Kutta formulae, their uses, and limitations*. J.Inst. Maths Applies, Vol.16, p.35-55. [II.5]
- C.F. Curtiss & J.O. Hirschfelder (1952): *Integration of stiff equations*. Proc. of the National Academy of Sciences of U.S., Vol.38, p.235-243. [III.1]
- G. Dahlquist (1956): *Convergence and stability in the numerical integration of ordinary differential equations*. Math. Scand., Vol.4, p.33-53. [III.2], [III.3], [III.4]
- G. Dahlquist (1959): *Stability and error bounds in the numerical integration of ordinary differential equations*. Trans. of the Royal Inst. of Techn., Stockholm, Sweden, Nr.130, 87pp. [I.10], [III.2], [III.10]
- G. Dahlquist (1985): *33 years of numerical instability, Part I*. BIT, Vol.25, p.188-204. [III.3]
- G. Dahlquist & R. Jeltsch (1979): *Generalized disks of contractivity for explicit and implicit Runge-Kutta methods*. Report TRITA-NA-7906, NADA, Roy. Inst. Techn. Stockholm. [II.12]
- G. Darboux (1876): *Sur les développements en série des fonctions d'une seule variable*. J. des Mathématiques pures et appl., 3ème série, t. II, p.291-312. [II.13]
- G. H. Darwin (Sir George) (1898): *Periodic orbits*. Acta Mathematica, Vol.21, p.99-242, plates I-IV. [II.0]
- S.M. Davenport, see L.F. Shampine, H.A. Watts & S.M. Davenport.
- F. Debaune (1638): *Letter to Descartes*. lost; answer of Descartes: Feb 20, 1639. [I.2]
- J.P. Den Hartog (1930): *Forced vibrations with combined viscous and Coulomb damping*. Phil. Mag. Ser.7, Vol.9, p.801-817. [II.6]
- J. Descoux (1963): *A note on a paper by A. Nordsieck*. Report No.131, Dept. of Comp. Sci., Univ. of Illinois at Urbana-Champaign. [III.6]
- P. Deuflhard (1980): *Recent advances in multiple shooting techniques*. In: Computational techniques for ordinary differential equations (Gladwell-Sayers, ed.), Section 10, p.217-272, Academic Press. [I.15]
- P. Deuflhard (1983): *Order and stepsize control in extrapolation methods*. Num. Math., Vol.41, p.399-422. [II.9], [II.10]
- P. Deuflhard (1985): *Recent progress in extrapolation methods for ordinary differential equations*. SIAM Rev., Vol.27, p.505-535. [II.14]
- P. Deuflhard & U. Nowak (1987): *Extrapolation integrators for quasilinear implicit ODEs*. In: P. Deuflhard, B. Engquist (eds.), Large-scale scientific computing, Birkhäuser, Boston. [II.9]
- E. de Doncker-Kapenga, see R. Piessens, E. de Doncker-Kapenga, C.W. Überhuber & D.K. Kahaner.
- J. Donelson & E. Hansen (1971): *Cyclic composite multistep predictor-corrector methods*. SIAM, J. Numer. Anal., Vol.8, p.137-157. [III.8]
- J.R. Dormand, M.E.A. El-Mikkawy & P.J. Prince (1987): *High-order embedded Runge-Kutta-Nystrom formulae*. IMA J. Numer. Anal., Vol.7, p.423-430. [II.14]

- J.R. Dormand & P.J. Prince (1978): *New Runge-Kutta algorithms for numerical simulation in dynamical astronomy*. Celestial Mechanics, Vol.18, p.223-232. [II.14]
- J.R. Dormand & P.J. Prince (1980): *A family of embedded Runge-Kutta formulae*. J.Comp. Appl. Math., Vol.6, p.19-26. [II.5]
- J.R. Dormand & P.J. Prince (1986): *Runge-Kutta triples*. Comp. & Maths. with Applc., Vol.12A, p.1007-1017. [II.6]
- J.R. Dormand & P.J. Prince (1987): *Runge-Kutta-Nystrom triples*. Comput. Math. Applic., Vol.13(12), p.937-949. [II.14]
- J.R. Dormand & P.J. Prince (1989): *Practical Runge-Kutta processes*. SIAM J. Sci. Stat. Comput., Vol.10, p.977-989. [II.5]
- J.R. Dormand, see also P.J. Prince & J.R. Dormand, R.W. Brankin, I. Gladwell, J.R. Dormand, P.J. Prince & W.L. Seward.
- R.D. Driver (1977): *Ordinary and delay differential equations*. Applied Math. Sciences 20, Springer Verlag, 501pp. [II.17]
- J.P. Eckmann, see P. Collet & J.P. Eckmann.
- B.L. Ehle (1968): *High order A-stable methods for the numerical solution of systems of D.E.'s*. BIT, Vol.8, p.276-278. [II.7]
- E. Eich (1992): *Projizierende Mehrschrittverfahren zur numerischen Lösung von Bewegungsgleichungen technischer Mehrkörpersysteme mit Zwangsbedingungen und Unstetigkeiten*. Fortschritt-Ber. VDI, Reihe 18, Nr.109, VDI-Verlag Düsseldorf, 188pp. [II.6]
- N.F. Eispack (1974): *B.T.Smith, J.M. Boyle, B.S.Garbow, Y.Jkebe, V.C.Klema, C.B.Moler: Matrix Eigensystem Routines*. (Fortran-translations of algorithms published in Reinsch & Wilkinson), Lecture Notes in Computer Science, Vol.6, Springer Verlag. [I.12], [I.13]
- M.E.A. El-Mikkawy, see J.R. Dormand, M.E.A. El-Mikkawy & P.J. Prince.
- H. Eltermann (1955): *Fehlerabschätzung bei näherungsweise Lösung von Systemen von Differentialgleichungen erster Ordnung*. Math. Zeitschr., Vol.62, p.469-501. [I.10]
- R. England (1969): *Error estimates for Runge-Kutta type solutions to systems of ordinary differential equations*. The Computer J. Vol.12, p.166-170. [II.4]
- W.H. Enright & D.J. Higham (1991): *Parallel defect control*. BIT, Vol.31, p.647-663. [II.11]
- W.H. Enright, K.R. Jackson, S.P. Nørsett & P.G. Thomson (1986): *Interpolants for Runge-Kutta formulas*. ACM Trans. Math. Softw., Vol.12, p.193-218. [II.6] [II.6]
- W.H. Enright, K.R. Jackson, S.P. Nørsett & P.G. Thomson (1988): *Effective solution of discontinuous IVPs using a Runge-Kutta formula pair with interpolants*. Appl. Math. and Comput., Vol.27, p.313-335. [II.6]
- T.O. Espelid, see J. Berntsen & T.O. Espelid.
- L. Euler (1728): *Nova methodus innumerabiles aequationes differentiales secundi gradus reducendi ad aequationes differentiales primi gradus*. Comm. acad. scient. Petrop., Vol.3, p.124-137; Opera Omnia, Vol.XXII, p.1-14. [I.3]
- L. Euler (1743): *De integratione aequationum differentialium altiorum graduum*. Miscellanea Berolinensia, Vol.7, p.193-242; Opera Omnia, Vol.XXII, p.108-149. See also: Letter from Euler to Joh. Bernoulli, 15.Sept.1739. [I.4]

- L. Euler (1744): *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes ...* Lausannae & Genevae, Opera Omnia (intr. by Caratheodory) Vol.XXIV, p.1-308. [I.2]
- L. Euler (1747): *Recherches sur le mouvement des corps celestes en général*. Hist. de l'Acad. Royale de Berlin, Tom.3, Année MDCCXLVII, publ. 1749, p.93-143. [I.6]
- L. Euler (1750): *Methodus aequationes differentiales altiorum graduum integrandi ulterius promota*. Novi Comment. acad. scient. Petrop., Vol.3, p.3-35; Opera Omnia, Vol.XXII, p.181-213. [I.4]
- L. Euler (1755): *Institutiones calculi differentialis cum eius vsu in analysi finitorum ac doctrina serierum*. Imp. Acad. Imper. Scient. Petropolitanae, Opera Omnia, Vol.X, [I.6]
- L. Euler (1756): *Elementa calculi variationum*. presented September 16, 1756 at the Acad. of Science, Berlin; printed 1766, Opera Omnia, Vol.XXV, p.141-176. [I.2]
- L. Euler (1758): *Du mouvement de rotation des corps solides autour d'un axe variable*. Hist. de l'Acad. Royale de Berlin, Tom.14, Année MDCCCLVIII, pp.154-193. Opera Omnia Ser.II, Vol.8, p.200-235. [II.10]
- L. Euler (1768): *Institutionum Calculi Integralis*. Volumen Primum, Opera Omnia, Vol.XI. [I.7], [I.8], [II.1]
- L. Euler (1769): *Institutionum Calculi Integralis*. Volumen Secundum, Opera Omnia, Vol.XII. [I.3], [I.5]
- L. Euler (1769b): *De formulis integralibus duplicatis*. Novi Comment. acad. scient. Petrop., Vol.14, I, 1770, p.72-103; Opera Omnia, Vol.XVII, p.289-315. [I.14]
- L. Euler (1778): *Specimen transformationis singularis serienum*. Nova acta. acad. Petrop., Vol.12 (1794), p.58-70, Opera Omnia, Vol.XVI, Sectio Altera, p.41-55. [I.5]
- E. Fehlberg (1958): *Eine Methode zur Fehlerverkleinerung beim Runge-Kutta-Verfahren*. ZAMM, Vol.38, p.421-426. [II.13]
- E. Fehlberg (1964): *New high-order Runge-Kutta formulas with step size control for systems of first and second order differential equations*. ZAMM, Vol.44, Sonderheft T17-T19. [II.4], [II.13]
- E. Fehlberg (1968): *Classical fifth-, sixth-, seventh-, and eighth order Runge-Kutta formulas with step size control*. NASA Technical Report 287 (1968); extract published in Computing, Vol.4, p.93-106 (1969). [II.4], [II.5]
- E. Fehlberg (1969): *Low-order classical Runge-Kutta formulas with step size control and their application to some heat transfer problems*. NASA Technical Report 315 (1969), extract published in Computing, Vol.6, p.61-71 (1970). [II.4], [II.5]
- E. Fehlberg (1972): *Classical eighth- and lower-order Runge-Kutta-Nyström formulas with stepsize control for special second-order differential equations*. NASA Technical Report R-381. [II.14]
- M. Feigenbaum (1978): *Quantitative universality for a class of nonlinear transformations*. J.Stat. Phys., Vol.19, p.25-52, Vol.21 (1979), p.669-706. [I.16]
- Feng Kang (冯康) (1985): *On difference schemes and symplectic geometry*. Proceedings of the 5-th Intern. Symposium on differential geometry & differential equations, August 1984, Beijing, p.42-58. [II.16]
- Feng Kang (1986): *Difference schemes for Hamiltonian formalism and symplectic geometry*. J. Comp. Math., Vol.4, p.279-289. [II.16]

- Feng Kang (1991): *How to compute properly Newton's equation of motion?* Proceedings of the second conference on numerical methods for partial differential equations, Nankai Univ., Tianjin, China, Eds. Ying Lungan & Guo Benyu, World Scientific, p.15-22. [II.16]
- Feng Kang (1991b): *Formal power series and numerical algorithms for dynamical systems.* Proceedings of international conference on scientific computation, Hangzhou, China, Eds. Tony Chan & Zhong-Ci Shi, Series on Appl. Math., Vol.1, pp.28-35. [II.16]
- Feng Kang, Wu Hua-mo, Qin Meng-zhao & Wang Dao-liu (1989): *Construction of canonical difference schemes for Hamiltonian formalism via generating functions.* J. Comp. Math., Vol.11, p.71-96. [II.16]
- J.R. Field & R.M. Noyes (1974): *Oscillations in chemical systems. IV. Limit cycle behavior in a model of a real chemical reaction.* J. Chem. Physics, Vol.60, p.1877-1884. [I.16]
- S. Filippi & J. Gräf (1986): *New Runge-Kutta-Nyström formula-pairs of order 8(7), 9(8), 10(9) and 11(10) for differential equations of the form $y'' = f(x, y)$.* J. Comput. and Applied Math., Vol.14, p.361-370. [II.14]
- A.F. Filippov (1960): *Differential equations with discontinuous right-hand side.* Mat. Sbornik (N.S.) Vol.51(93), p.99-128; Amer. Math. Soc. Transl. Ser.2, Vol.42, p.199-231. [II.6]
- J.M. Fine (1987): *Interpolants for Runge-Kutta-Nyström methods.* Computing, Vol.39, p.27-42. [II.14]
- R. Fletcher & D.C. Sorensen (1983): *An algorithmic derivation of the Jordan canonical form.* Amer. Math. Monthly, Vol.90, No.1, p.12-16. [I.12]
- C.V.D. Forrrington (1961-62): *Extensions of the predictor-corrector method for the solution of systems of ordinary differential equations.* Comput. J. 4, p.80-84. [III.5]
- J.B.J. Fourier (1807): *Sur la propagation de la chaleur.* Unpublished manuscript; published: La théorie analytique de la chaleur, Paris 1822. [I.6]
- R.A. Frazer, W.P. Jones & S.W. Skan (1937): *Approximations to functions and to the solutions of differential equations.* Reports and Memoranda Nr.1799 (2913), Aeronautical Research Committee. 33pp. [II.7]
- A. Fricke (1949): *Ueber die Fehlerabschätzung des Adamsschen Verfahrens zur Integration gewöhnlicher Differentialgleichungen erster Ordnung.* ZAMM, Vol.29, p.165-178. [III.4]
- G. Frobenius (1873): *Ueber die Integration der linearen Differentialgleichungen durch Reihen.* Journal für Math. LXXVI, p.214-235 [I.5]
- M. Frommer (1934): *Ueber das Auftreten von Wirbeln und Strudeln (geschlossener und spiraliger Integralkurven) in der Umgebung rationaler Unbestimmtheitsstellen.* Math. Ann., Vol.109, p.395-424. [I.16]
- L. Fuchs (1866, 68): *Zur Theorie der linearen Differentialgleichungen mit veränderlichen Coefficienten.* Crelle J. f. d. r. u. angew. Math., Vol.66, p.121-160 (prepublished in "Programm der städtischen Gewerbeschule zu Berlin, Ostern 1865"). Ergänzung: J. f. Math. LXVIII, p. 354-385. [I.5], [I.11]
- C.F. Gauss (1812): *Disquisitiones generales circa seriem infinitam*

$$1 + \frac{\alpha\beta}{1\cdot\gamma}x + \frac{\alpha(\alpha+1)\beta(\beta+1)}{1\cdot2\cdot\gamma(\gamma+1)}xx + \frac{\alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)}{1\cdot2\cdot3\cdot\gamma(\gamma+1)(\gamma+2)}x^3 + \text{etc},$$
Werke, Vol.3, p.123-162. [I.5]

- W. Gautschi (1962): *On inverses of Vandermonde and confluent Vandermonde matrices*. Numer. Math., Vol.4, p.117-123. [II.13]
- C.W. Gear (1965): *Hybrid methods for initial value problems in ordinary differential equations*. SIAM J. Numer. Anal., ser.B, Vol.2, p.69-86. [III.8]
- C.W. Gear (1971): *Numerical initial value problems in ordinary differential equations*. Prentice-Hall, 253pp. [II.2], [III.1], [III.7]
- C.W. Gear (1987): *The potential for parallelism in ordinary differential equations*. In: Computational mathematics II, Proc. 2nd Int. Conf. Numer. Anal. Appl., Benin City/Niger. 1986, Conf. Ser. Boole Press 11, p. 33-48. [II.11]
- C.W. Gear (1988): *Parallel methods for ordinary differential equations*. Calcolo, Vol.25, No.1/2, p. 1-20. [II.11]
- C.W. Gear & O. Østerby (1984): *Solving ordinary differential equations with discontinuities*. ACM Trans. Math. Softw., Vol.10, p.23-44. [II.6]
- C.W. Gear & K.W. Tu (1974): *The effect of variable mesh size on the stability of multistep methods*. SIAM J. Num. Anal., Vol.11, p.1025-1043. [III.5]
- C.W. Gear & D.S. Watanabe (1974): *Stability and convergence of variable order multistep methods*. SIAM J. Num. Anal., Vol.11, p.1044-1058. [III.3]
- W.J. Gerritsen, see P.A. Beentjes & W.J. Gerritsen.
- A. Gibbons (1960): *A program for the automatic integration of differential equations using the method of Taylor series*. Computer J., Vol.3, p.108-111. [I.8]
- S. Gill (1951): *A process for the step-by-step integration of differential equations in an automatic digital computing machine*. Proc. Cambridge Philos. Soc., Vol.47, p.95-108. [II.1], [II.2]
- S. Gill (1956): Discussion in Merson (1957). [II.2]
- B. Giovannini, L. Weiss-Parmeggiani & B.T. Ulrich (1978): *Phase locking in coupled Josephson weak links*. Helvet. Physica Acta, Vol.51, p.69-74. [I.16]
- I. Gladwell, L.F. Shampine & R.W. Brankin (1987): *Automatic selection of the initial step size for an ODE solver*. J. Comp. Appl. Math., Vol.18, p.175-192. [II.4]
- I. Gladwell, see also R.W. Brankin, I. Gladwell, J.R. Dormand, P.J. Prince & W.L. Seward.
- G.H. Golub & J.H. Wilkinson (1976): *Ill-conditioned eigensystems and the computation of the Jordan canonical form*. SIAM Review, Vol.18, p.578-619. [I.12]
- M.K. Gordon, see L.F. Shampine & M.K. Gordon.
- J. Gräf, see S. Filippi & J. Gräf.
- W.B. Gragg (1964): *Repeated extrapolation to the limit in the numerical solution of ordinary differential equations*. Thesis, Univ. of California; see also SIAM J. Numer. Anal., Vol.2, p.384-403 (1965). [II.8], [II.9]
- W.B. Gragg (1965): *On extrapolation algorithms for ordinary initial value problems*. SIAM J. Num. Anal., ser.B, Vol.2, p.384-403. [II.14]
- W.B. Gragg & H.J. Stetter (1964): *Generalized multistep predictor-corrector methods*. J. ACM, Vol.11, p.188-209. [III.8]

- E. Griepentrog (1978): *Gemischte Runge-Kutta-Verfahren für steife Systeme*. In: Seminarbericht Nr. 11, Sekt. Math., Humboldt-Univ. Berlin, p.19-29. [II.15]
- R.D. Grigorieff (1977): *Numerik gewöhnlicher Differentialgleichungen 2*. Teubner Studienbücher, Stuttgart. [III.3], [III.4]
- R.D. Grigorieff (1983): *Stability of multistep-methods on variable grids*. Numer. Math. 42, p.359-377. [III.5]
- W. Gröbner (1960): *Die Liereihen und ihre Anwendungen*. VEB Deutscher Verlag der Wiss., Berlin 1960, 2nd ed. 1967. [I.14], [II.16]
- T.H. Gronwall (1919): *Note on the derivatives with respect to a parameter of the solutions of a system of differential equations*. Ann. Math., Vol.20, p.292-296. [I.10], [I.14]
- A. Guillou & J.L. Soulé (1969): *La résolution numérique des problèmes différentiels aux conditions initiales par des méthodes de collocation*. R.I.R.O, No R-3, p.17-44. [II.7]
- P. Habets, see N. Rouche, P. Habets & M. Laloy.
- H. Hahn (1921): *Theorie der reellen Funktionen*. Springer Verlag Berlin, 600pp. [I.7]
- W. Hahn (1967): *Stability of motion*. Springer Verlag, 446pp. [I.13]
- E. Hairer (1977): *Méthodes de Nyström pour l'équation différentielle $y'' = f(x, y)$* . Numer. Math., Vol.27, p.283-300. [II.14]
- E. Hairer (1978): *A Runge-Kutta method of order 10*. J.Inst. Maths Applies, Vol.21, p.47-59. [II.5]
- E. Hairer (1981): *Order conditions for numerical methods for partitioned ordinary differential equations*. Numer. Math., Vol.36, p.431-445. [II.15]
- E. Hairer (1982): *A one-step method of order 10 for $y'' = f(x, y)$* . IMA J. Num. Anal., Vol.2, p.83-94. [II.14]
- E. Hairer & Ch. Lubich (1984): *Asymptotic expansions of the global error of fixed-stepsize methods*. Numer. Math., Vol.45, p.345-360. [II.8], [III.9]
- E. Hairer & A. Ostermann (1990): *Dense output for extrapolation methods*. Numer. Math., Vol.58, p.419-439. [III.9]
- E. Hairer & A. Ostermann (1992): *Dense output for the GBS extrapolation method*. In: Computational ordinary differential equations, ed. by J.R. Cash & I. Gladwell, Clarendon Press, Oxford, p.107-114. [II.9]
- E. Hairer & G. Wanner (1973): *Multistep-multistage-multiderivative methods for ordinary differential equations*. Computing, Vol.11, p.287-303. [II.13], [III.8]
- E. Hairer & G. Wanner (1974): *On the Butcher group and general multi-value methods*. Computing, Vol.13, p.1-15. [II.2], [II.12], [II.13]
- E. Hairer & G. Wanner (1976): *A theory for Nyström methods*. Numer. Math., Vol.25, p.383-400. [II.14]
- E. Hairer & G. Wanner (1983): *On the instability of the BDF formulas*. SIAM J. Numer. Anal., Vol.20, No.6, p.1206-1209. [III.3]
- Sir W. R. Hamilton (1833): *On a general method of expressing the paths of light, and of the planets, by the coefficients of a characteristic function*. Dublin Univ. Review, p.795-826; Math. Papers, Vol.I, p.311-332. [I.6]

- Sir W. R. Hamilton (1834): *On a general method in dynamics; by which the study of the motions of all free systems of attracting or repelling points is reduced to the search and differentiation of one central relation, or characteristic function*. Phil. Trans. Roy. Soc. Part II for 1834, p.247-308; Math. Papers, Vol.II, p.103-161. [I.6]
- Sir W. R. Hamilton (1835): *Second essay on a general method in dynamics*. Phil. Trans. Roy. Soc. Part I for 1835, p.95-144; Math. Papers, Vol.II, p.162-211. [I.6]
- P.C. Hammer & J.W. Hollingsworth (1955): *Trapezoidal methods of approximating solutions of differential equations*. MTAC, Vol.9, p.92-96. [II.7]
- E. Hansen, see J. Donelson & E. Hansen.
- F. Hausdorff (1906): *Die symbolische Exponentialformel in der Gruppentheorie*. Berichte ü. d. Verh. Königl. Sächs. Ges. d. Wiss. Leipzig, Math.-Phys. Klasse, Vol.58, p.19-48. [II.16]
- K. Hayashi, see M. Okamoto & K. Hayashi.
- N.D. Hayes (1950): *Roots of the transcendental equation associated with a certain difference-differential equation*. J. of London Math. Soc., Vol.25, p.226-232. [II.17]
- H.M. Hebsacker (1982): *Conditions for the coefficients of Runge-Kutta methods for systems of n -th order differential equations*. J. Comput. Appl. Math., Vol.8, p.3-14. [II.14]
- P. Henrici (1962): *Discrete variable methods in ordinary differential equations*. John Wiley & Sons, Inc., New-York-London-Sydney. [II.2], [II.8], [III.1], [III.2], [III.10]
- P. Henrici (1974): *Applied and computational complex analysis*. Volume 1, John Wiley & Sons, New York, 682pp. [I.13], [III.3]
- Ch. Hermite (1878): *Extrait d'une lettre de M. Ch. Hermite à M. Borchardt sur la formule d'interpolation de Lagrange*. J. de Crelle, Vol.84, p.70; Oeuvres, tome III, p.432-443. [II.13]
- K. Heun (1900): *Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen*. Zeitschr. für Math. u. Phys., Vol.45, p.23-38. [I.5], [II.1]
- D.J. Higham, see W.H. Enright & D.J. Higham.
- A.C. Hindmarsh (1972): *GEAR: ordinary differential equation system solver*. UCID-30001, Rev.2, LLL, Livermore, Calif. [III.7]
- A.C. Hindmarsh (1980): *LSODE and LSODI, two new initial value ordinary differential equation solvers*. ACM Signum Newsletter 15,4. [II.4]
- A.C. Hindmarsh, see also P.N. Brown, G.D. Byrne & A.C. Hindmarsh, G.D. Byrne & A.C. Hindmarsh.
- J.O. Hirschfelder, see C.F. Curtiss & J.O. Hirschfelder.
- E.W. Hobson (1921): *The theory of functions of a real variable*. Vol.I, Cambridge, 670pp. [I.10]
- E. Hofer (1976): *A partially implicit method for large stiff systems of ODEs with only few equations introducing small time-constants*. SIAM J. Numer. Anal., vol 13, No.5, p.645-663. [II.15]
- J.W. Hollingsworth, see P.C. Hammer & J.W. Hollingsworth.

- G.'t Hooft (1974): *Magnetic monopoles in unified gauge theories*. Nucl. Phys., Vol.B79, p.276-284. [I.6]
- E. Hopf (1942): *Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems*. Ber. math. physik. Kl. Akad. d. Wiss. Leipzig, Bd.XCIV, p.3-22. [I.16]
- M.K. Horn (1983): *Fourth and fifth-order scaled Runge-Kutta algorithms for treating dense output*. SIAM J.Numer. Anal., Vol.20, p.558-568. [II.6]
- P.J. van der Houwen (1977): *Construction of integration formulas for initial value problems*. North-Holland Amsterdam, 269pp. [II.1]
- P.J. van der Houwen & B.P. Sommeijer (1990): *Parallel iteration of high-order Runge-Kutta methods with step size control*. J. Comput. Appl. Math., Vol.29, p.111-127. [II.11]
- P.J. van der Houwen, see also H. Brunner & P.J. van der Houwen.
- T.E. Hull (1967): A search for optimum methods for the numerical integration of ordinary differential equations. SIAM Rev., Vol.9, p.647-654. [II.1], [II.3]
- T.E. Hull & R.L. Johnston (1964): *Optimum Runge-Kutta methods*. Math. Comput., Vol.18, p.306-310. [II.3]
- B.L. Hulme (1972): *One-step piecewise polynomial Galerkin methods for initial value problems*. Math. of Comput., Vol.26, p.415-426. [II.7]
- W.H. Hundsdorfer & M.N. Spijker (1981): *A note on B-stability of Runge-Kutta methods*. Num. Math., Vol.36, p.319-331. [II.12]
- A. Hurwitz (1895): *Ueber die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Theilen besitzt*. Math. Ann., Vol.46, p.273-284; Werke, Vol.2, p.533ff. [I.13]
- A. Huťa (1956): *Une amélioration de la méthode de Runge-Kutta-Nyström pour la résolution numérique des équations différentielles du premier ordre*. Acta Fac. Rerum Natur. Univ. Comenianae (Bratislava) Math., Vol.1, p.201-224. [II.5]
- B.J. Hyett, see R.F. Warming & B.J. Hyett.
- E.L. Ince (1944): *Ordinary differential equations*. Dover Publications, New York, 558pp. [I.2], [I.3]
- A. Iserles & S.P. Nørsett (1990): On the theory of parallel Runge-Kutta methods. IMA J. Numer. Anal., Vol.10, p.463-488. [II.11]
- Z. Jackiewicz & M. Zennaro (1992): *Variable stepsize explicit two-step Runge-Kutta methods*. Math. Comput., Vol.59, p.421-438. [III.8]
- K.R. Jackson (1991): *A survey of parallel numerical methods for initial value problems for ordinary differential equations*. IEEE Trans. on Magnetics, Vol.27, p.3792-3797. [II.11]
- K.R. Jackson & S.P. Nørsett (1992): *The potential of parallelism in Runge-Kutta methods. Part I: RK formulas in standard form*. Report. [II.11]
- K.R. Jackson, see also W.H. Enright, K.R. Jackson, S.P. Nørsett & P.G. Thomson.
- C.G.J. Jacobi (1841): *De determinantibus functionalibus*. Crelle J. f. d. r. u. angew. Math., Vol.22, p.319-359, Werke, Vol.III, p.393-438. [I.14]

- C.G.J. Jacobi (1842/43): *Vorlesungen über Dynamik*, gehalten an der Universität zu Königsberg im Wintersemester 1842–1843 und nach einem von C.W. Borchardt ausgearbeiteten Hefte, edited 1866 by A. Clebsch, Werke, Vol. VIII. [I.6],[I.14]
- C.G.J. Jacobi (1845): *Theoria novi multiplicatoris systemati aequationum differentialium vulgarium applicandi*. Crelle J. f. d. r. u. angew. Math, Vol.29, p.213-279, 333-376. Werke, Vol.IV, p.395-509. [I.11]
- R. Jeltsch, see G. Dahlquist & R. Jeltsch.
- R.L. Johnston, see T.E. Hull & R.L. Johnston.
- W.P. Jones, see R.A. Frazer, W.P. Jones & S.W. Skan.
- C. Jordan (1870): *Traité des Substitutions et des équations algébriques*. Paris 667pp. [I.12]
- C. Jordan (1928): *Sur une formule d'interpolation*. Atti Congresso Bologna, vol 6, p.157-177 [II.9]
- B. Kågström & A. Ruhe (1980): *An algorithm for numerical computation of the Jordan normal form of a complex matrix*. ACM Trans. Math. Software, Vol.6, p.398-419. (Received May 1975, revised Aug. 1977, accepted May 1979). [I.12]
- D.K. Kahaner, see R. Piessens, E. de Doncker-Kapenga, C.W. Überhuber & D.K. Kahaner.
- S. Kakutani & L. Marcus (1958): *On the non-linear difference-differential equation $y'(t) = [A - By(t - \tau)]y(t)$* . In: Contributions to the theory of nonlinear oscillations, ed. by S.Lefschetz, Princeton, Vol.IV, p.1-18. [II.17]
- E. Kamke (1930): *Ueber die eindeutige Bestimmtheit der Integrale von Differentialgleichungen II*. Sitz. Ber. Heidelberg Akad. Wiss. Math. Naturw. Kl., 17. Abhandl., see also Math. Zeitschr., Vol.32, p.101-107. [I.10]
- E. Kamke (1942): *Differentialgleichungen, Lösungsmethoden und Lösungen*. Becker & Erler, Leipzig, 642pp. [I.3]
- K.H. Kastlunger & G. Wanner (1972): *Runge Kutta processes with multiple nodes*. Computing, Vol.9, p.9-24. [II.13]
- K.H. Kastlunger & G. Wanner (1972b): *On Turan type implicit Runge-Kutta methods*. Computing, Vol.9, p.317-325. [II.13]
- A.G.Mc. Kendrick, see W.O. Kermack & A.G.Mc. Kendrick.
- W.O. Kermack & A.G.Mc. Kendrick (1927): *Contributions to the mathematical theory of epidemics (Part I)*. Proc. Roy. Soc., A, Vol.115, p.700-721. [II.17]
- H. Knapp & G. Wanner (1969): *LIESE II, A program for ordinary differential equations using Lie-series*. MRC Report No.1008, Math. Research Center, Univ. Wisconsin, Madison, Wisc. 53706. [I.8]
- H. König, see C. Runge & H. König.
- F.T. Krogh (1969): *A variable step variable order multistep method for the numerical solution of ordinary differential equations*. Information Processing 68, North-Holland, Amsterdam, p.194-199. [III.5]
- F.T. Krogh (1973): *Algorithms for changing the step size*. SIAM J. Num. Anal. 10, p.949-965. [III.5]

- F.T. Krogh (1974): *Changing step size in the integration of differential equations using modified divided differences*. Proceedings of the Conference on the Num. Sol. of ODE, Lecture Notes in Math. No.362, Springer Verlag New York, p.22-71. [III.5]
- N. Kryloff & N. Bogoliuboff (1947): *Introduction to non-linear Mechanics*. Free translation by S. Lefschetz, Princeton Univ. Press, 105pp. [I.16]
- E.E. Kummer (1839): *Note sur l'intégration de l'équation $d^m y/dx^n = x^m y$ par des intégrales définies*. Crelle J. f. d. r. u. angew. Math., Vol.19, p.286-288. [I.11]
- J. Kuntzmann (1961): *Neuere Entwicklungen der Methode von Runge-Kutta*. ZAMM, Vol.41, p.28-31. [II.7]
- J. Kuntzmann, see also F. Ceschino & J. Kuntzmann.
- W. Kutta (1901): *Beitrag zur näherungsweise Integration totaler Differentialgleichungen*. Zeitschr. für Math. u. Phys., Vol.46, p.435-453. [II.1], [II.2], [II.3], [II.5]
- J.L.de Lagrange (1759): *Recherches sur la nature et la propagation du son*. Miscell. Taurinensia t.I, Oeuvres t.1, p.39-148. [I.6]
- J.L.de Lagrange (1762): *Solution de différents problèmes de Calcul Intégral*. Miscell. Taurinensia, t.III, Oeuvres t.1, p.471-668. [I.6]
- J.L.de Lagrange (1774): *Sur les Intégrales particulières des Equations différentielles*. Oeuvres, tom.4, p.5-108. [I.2]
- J.L.de Lagrange (1775): *Recherche sur les Suites Récurentes*. Nouveaux Mém. de l'Acad. royale des Sciences et Belles-Lettres, Berlin. Oeuvres, Vol.4, p.159. [I.4], [III.3]
- J.L.de Lagrange (1788): *Mécanique analytique*. Paris, Oeuvres t.11 et 12. [I.4], [I.6], [I.12]
- J.L.de Lagrange (1792): *Mémoire sur l'expression du terme général des séries récurrentes, lorsque l'équation génératrice a des racines égales*. Nouv. Mém. de l'Acad. royale des Sciences de Berlin, Oeuvres t.5, p.627-641. [III.3]
- J.L.de Lagrange (1797): *Théorie des fonctions analytiques, contenant les principes du calcul différentiel, dégagés de toute considération d'infiniment petits, d'évanouissants, de limites et de fluxions, et réduits à l'analyse algébrique des quantités finies*. Paris, 1797, nouv. ed. 1813, Oeuvres Tome 9. [II.3]
- E. Laguerre (1883): *Mémoire sur la théorie des équations numériques*. J. Math. pures appl. (3e série), Vol.9, p.99-146 (also in *Oeuvres I*, p.3-47). [II.9]
- J.D. Lambert (1987): *Developments in stability theory for ordinary differential equations*. In: The state of the art in numerical analysis, ed. by A. Iserles & M.J.D. Powell, Clarendon Press, Oxford, p.409-431. [I.13]
- M. Laloy, see N. Rouche, P. Habets & M. Laloy.
- R.J. Lambert, see G.D. Byrne & R.J. Lambert.
- P.S. Laplace (An XIII = 1805): *Supplément au dixième livre du Traité de mécanique céleste sur l'action capillaire*. Paris chez Courcier, 65+78pp. [II.1]
- F.M. Lasagni (1988): *Canonical Runge-Kutta methods*. ZAMP Vol.39, p.952-953. [II.16]
- F.M. Lasagni (1990): *Integration methods for Hamiltonian differential equations*. Unpublished manuscript. [II.16]

- P.D. Lax & R.D. Richtmyer (1956): *Survey of the stability of linear limite difference equations*. Comm. Pure Appl. Math., Vol.9, p.267-293. [III.4]
- R. Lefever & G. Nicolis (1971): *Chemical Instabilities and sustained oscillations*. J. theor. Biol., Vol.30, p.267-284. [I.16]
- A.M. Legendre (1787): *Mémoire sur l'intégration de quelques équations aux différences partielles*. Histoire Acad. R. Sciences, Paris, Année MDCCLXXXVII, à Paris MDCCLXXXIX, p.309-351. [I.6]
- G.W. Leibniz (1684): *Nova methodus pro maximis et minimis, itemque tangentibus, quae nec fractas, nec irrationales quantitates moratur, & singulare pro illis calculi genus*. Acta Eruditorum, Lipsiae, MDCLXXXIV, p.467-473. [I.2]
- G.W. Leibniz (1691): *Methodus, qua innummerarum linearum construction ex data proprietate tangentium seu aequatio inter abscissam et ordinatam ex dato valore subtangentialis, exhibetur* Letter to Huygens, in: C.I. Gerhardt, Leibnizens math. Schriften, 1850, Band II, p.116-121. [I.3]
- G.W. Leibniz (1693) (Gothofredi Guilielmi Leibnitzii): *Supplementum Geometriae Dimensoriae seugeneralissima omnium tetra gonismorum effectio per motum: Similiterque multiplex constructio lineae ex data tangentium conditione*. Acta Eruditorum, Lipsiae, p.385-392; german translation: G. Kowalewski, Leibniz über die Analysis des Unendlichen, Ostwalds Klassiker Nr.162 (1908), p.24-34. [I.2]
- A.M. Liapunov (1892): *Problème général de la stabilité du mouvement*. Russ., trad. en français 1907 (Annales de la Faculté des Sciences de Toulouse), reprinted 1947 Princeton Univ. Press, 474pp. [I.13]
- A.M. Liénard (1928): *Etude des oscillations entretenues*. Revue générale de l'Electricité, tome XXIII, p. 901-912 et 946-954. [I.16]
- B. Lindberg (1972): *A simple interpolation algorithm for improvement of the numerical solution of a differential equation*. SIAM J. Numer. Anal., Vol.9, p.662-668. [II.9]
- E. Lindelöf (1894): *Sur l'application des méthodes d'approximation successives à l'étude des intégrales réelles des équations différentielles ordinaires*. J. de Math., 4e série, Vol.10, p.117-128. [I.8]
- W. Liniger, see W.L. Miranker & W. Liniger.
- J. Liouville (1836): *Sur le développement des fonctions ou parties de fonctions en séries dont les divers termes sont assujétis à satisfaire à une même équation différentielle du second ordre, contenant un paramètre variable*. Journ. de Math. pures et appl., Vol.1, p.253-265. [I.8], [I.15]
- J. Liouville (1838): *Sur la Théorie de la variation des constantes arbitraires*. Liouville J. de Math., Vol.3, p.342-349. [I.8], [I.11]
- J. Liouville (1841): *Remarques nouvelles sur l'équation de Riccati*. J. des Math. pures et appl., Vol.6, p.1-13. [I.3]
- R. Lipschitz (1876): *Sur la possibilité d'intégrer complètement un système donné d'équations différentielles*. Bulletin des Sciences Math. et Astr., Paris, Vol.10, p.149-159. [I.7]
- F.J. Lisbona, see M. Crouzeix & F.J. Lisbona.
- R. Lobatto (1852): *Lessen over Differentiaal- en Integraal-Rekening*. 2 Vol., La Haye 1851-52. [II.7]

- E.N. Lorenz (1979): *On the prevalence of aperiodicity in simple systems*. Global Analysis, Calgary 1978, ed. by M.Grmela and J.E.Marsden, Lecture Notes in Mathematics, Vol.755, p.53-75. [I.16]
- F.R. Loscalzo (1969): *An introduction to the application of spline functions to initial value problems*. In: Theory and Applications of spline functions, ed. T.N.E. Greville, Acad. Press 1969, p.37-64. [II.13]
- F.R. Loscalzo & I.J. Schoenberg (1967): *On the use of spline functions for the approximation of solutions of ordinary differential equations*. Tech. Summ. Rep. # 723, Math. Res. Center, Univ. Wisconsin, Madison. [II.13]
- M. Lotkin (1951): *On the accuracy of Runge-Kutta methods*. MTAC Vol.5, p.128-132. [II.3]
- Ch. Lubich (1989): *Linearly implicit extrapolation methods for differential-algebraic systems*. Numer. Math., Vol.55, p.197-211. [II.9]
- Ch. Lubich, see also E. Hairer & Ch. Lubich.
- G.I. Marchuk (1975): *Prostejshaya matematicheskaya model virusnogo zabolevaniya*. Novosibirsk, VTS SO AN SSSR. Preprint. [II.17]
- G.I. Marchuk (1983): *Mathematical models in immunology*. Translation series, Optimization Software, New York, Springer Verlag, 351pp. [II.17]
- L. Marcus, see S. Kakutani & L. Marcus.
- M. Marden (1966): *Geometry of polynomials*. American Mathematical Society, Providence, Rhode Island, 2nd edition. [III.3]
- R.M. May (1976): *Simple mathematical models with very complicated dynamics*. Nature, Vol.261, p.459-467 [I.16]
- R.M.M. Mattheij, see U.M. Ascher, R.M.M. Mattheij & R.D. Russel.
- R.H. Merson (1957): *An operational method for the study of integration processes*. Proc. Symp. Data Processing, Weapons Research Establishment, Salisbury, Australia, p.110-1 to 110-25. [II.2], [II.4], [II.14]
- S. Miesbach & H.J. Pesch (1992): *Symplectic phase flow approximation for the numerical integration of canonical systems*. Numer.Math., Vol.61, p.501-521. [II.16]
- J.J.H. Miller, see D.M. Creedon & J.J.H. Miller.
- W.E. Milne (1926): *Numerical integration of ordinary differential equations*. Amer. Math. Monthly, Vol.33, p.455-460. [III.1]
- W.E. Milne (1970): *Numerical solution of differential equations*. Dover Publications, Inc., New York, second edition. [III.1]
- W.L. Miranker (1971): *A survey of parallelism in numerical analysis*. SIAM Review, Vol.13, p.524-547. [II.11]
- W.L. Miranker & W. Liniger (1967): *Parallel methods for the numerical integration of ordinary differential equations*. Math. Comput., Vol.21, p. 303-320. [II.11]
- R. von Mises (1930): *Zur numerischen Integration von Differentialgleichungen*. ZAMM, Vol.10, p.81-92. [III.4]

- A.R. Mitchell & J.W. Craggs (1953): *Stability of difference relations in the solution of ordinary differential equations*. Math. Tables Aids Comput., Vol.7, p.127-129. [III.1], [III.3]
- C. Moler & C. Van Loan (1978): *Nineteen dubious ways to compute the exponential of a matrix*; SIAM Review, Vol.20, p.801-836. [I.12]
- J.I. Montijano, see M. Calvo, J.I. Montijano & L. Rández.
- R.E. Moore (1966): *Interval Analysis*. Prentice-Hall, Inc, 145pp. [I.8]
- R.E. Moore (1979): *Methods and applications of interval analysis*. SIAM studies in Appl. Math., 190pp. [I.8]
- P. Moss, see K. Burrage & P. Moss.
- F.R. Moulton (1926): *New methods in exterior ballistics*. Univ. Chicago Press. [III.1]
- M. Müller (1926): *Ueber das Fundamentaltheorem in der Theorie der gewöhnlichen Differentialgleichungen*. Math. Zeitschr., Vol.26, p.619-645. (Kap.III). [I.10]
- F.D. Murnaghan, see A. Wintner & F.D. Murnaghan.
- O. Nevanlinna (1989): *Remarks on Picard-Lindelöf iteration*. BIT, Vol.29, p.328-346 and 535-562. [I.8]
- E.H. Neville (1934): *Iterative interpolation*. Ind. Math. Soc. J. Vol.20, p.87-120. [II.9]
- I. Newton (1671): *Methodus Fluxionum et Serierum Infinitarum*. edita Londini 1736, Opuscula mathematica, Vol.I, Traduit en français par M.de Buffon, Paris MDCCXL. [I.2]
- I. Newton (1687): *Philosophiae naturalis principia mathematica*. Imprimatur S. Pepys, Reg. Soc. Praeses, julii 5, 1686, Londini anno MDCLXXXVII. [I.6], [II.14]
- I. Newton (1711): *Methodus differentialis (Analysis per quantitatum, series, fluxiones, ac differentias: cum enumeratione linearum tertii ordinis)*. London 1711. [III.1]
- G. Nicolis, see R. Lefever & G. Nicolis.
- J. Nievergelt (1964): *Parallel methods for integrating ordinary differential equations*. Comm. ACM, Vol.7, p.731-733. [II.11]
- S.P. Nørsett (1974a): *One-step methods of Hermite type for numerical integration of stiff systems*. BIT, Vol.14, p.63-77. [II.13]
- S.P. Nørsett (1974b): *Semi explicit Runge-Kutta methods*. Report No.6/74, ISBN 82-7151-009-6, Dept. Math. Univ. Trondheim, Norway, 68+7pp. [II.7]
- S.P. Nørsett & G. Wanner (1981): *Perturbed collocation and Runge-Kutta methods*. Numer. Math., Vol.38, p.193-208. [II.7]
- S.P. Nørsett, see also A. Iserles & S.P. Nørsett, K.R. Jackson & S.P. Nørsett, W.H. Enright, K.R. Jackson, S.P. Nørsett & P.G. Thomson.
- A. Nordsieck (1962): *On numerical integration of ordinary differential equations*. Math. Comp., Vol.16, p.22-49. [III.6]
- U. Nowak, see P. Deuffhard & U. Nowak.
- R.M. Noyes, see J.R. Field & R.M. Noyes.
- B. Numerov (B.V.Noumerov) (1924): *A method of extrapolation of perturbations*. Monthly notices of the Royal Astronomical Society, Vol.84, p.592-601. [III.10]

- B. Numerov (1927): *Note on the numerical integration of $d^2x/dt^2 = f(x, t)$* . Astron. Nachrichten, Vol.230, p.359-364. [III.10]
- E.J. Nyström (1925): *Ueber die numerische Integration von Differentialgleichungen*. Acta Soc. Sci. Fenn., Vol.50, No.13, p.1-54. [II.2], [II.14], [III.1]
- H.J. Oberle & H.J. Pesch (1981): *Numerical treatment of delay differential equations by Hermite interpolation*. Numer. Math., Vol.37, p.235-255. [II.17]
- N. Obreschkoff (1940): *Neue Quadraturformeln*. Abh. der Preuss. Akad. der Wiss., Math.-naturwiss. Klasse, Nr.4, Berlin. [II.13]
- M. Okamoto & K. Hayashi (1984): *Frequency conversion mechanism in enzymatic feedback systems*. J. Theor. Biol., Vol.108, p.529-537. [II.17]
- D. Okunbor & R.D. Skeel (1992): *An explicit Runge-Kutta-Nyström method is canonical if and only if its adjoint is explicit*. SIAM J. Numer. Anal., Vol.29, p. 521-527. [II.16]
- D. Okunbor & R.D. Skeel (1992b): *Explicit canonical methods for Hamiltonian systems*. Math. Comput., Vol.59, p.439-455. [II.16]
- D. Okunbor & R.D. Skeel (1992c): *Canonical Runge-Kutta-Nyström methods of orders 5 and 6*, Working Document 92-1, Dep. Computer Science, Univ. Illinois. [II.16]
- J. Oliver (1975): *A curiosity of low-order explicit Runge-Kutta methods*. Math. Comp., Vol.29, p.1032-1036. [II.1]
- J. Ooppelstrup (1976): *The RKFHB4 method for delay-differential equations*. Lect. Notes Math., Nr. 631, p.133-146. [II.17]
- M.R. Osborne (1966): *On Nordsieck's method for the numerical solution of ordinary differential equations*. BIT, Vol.6, p.51-57. [III.6]
- O. Østerby, see C.W. Gear & O. Østerby.
- A. Ostermann, see E. Hairer & A. Ostermann.
- B. Owren & M. Zennaro (1991): *Order barriers for continuous explicit Runge-Kutta methods*. Math. Comput., Vol.56, p.645-661. [II.6]
- B. Owren & M. Zennaro (1992): *Derivation of efficient, continuous, explicit Runge-Kutta methods*. SIAM J. Sci. Stat. Comput., Vol.13, p.1488-1501. [II.6]
- B.N. Parlett (1976): *A recurrence among the elements of functions of triangular matrices*. Linear Algebra Appl., Vol.14, p.117-121. [I.12]
- G. Peano (1888): *Intégration par séries des équations différentielles linéaires*. Math. Annalen, Vol.32, p.450-456. [I.8], [I.9]
- G. Peano (1890): *Démonstration de l'intégrabilité des équations différentielles ordinaires*, Math. Annalen, Vol.37, p.182-228; see also the german translation and commentation: G. Mie, Math. Annalen, Vol.43 (1893), p.553-568. [I.1], [I.7], [I.9], [I.10]
- G. Peano (1913): *Resto nelle formule di quadratura, espresso con un integrale definito*. Atti Della Reale Accad. Dei Lincei, Rendiconti, Vol.22, N.9, p.562-569, Roma. [III.2]
- R. Pearl & L.J. Reed (1922): *A further note on the mathematical theory of population growth*. Proceedings of the National Acad. of Sciences, Vol.8, No.12, p.365-368. [II.17]
- L.M. Perko (1984): *Limit cycles of quadratic systems in the plane*. Rocky Mountain J. of Math., Vol.14, p.619-645. [I.16]

- O. Perron (1915): *Ein neuer Existenzbeweis für die Integrale der Differentialgleichung $y' = f(x, y)$* . Math. Annalen, Vol.76, p.471-484. [I.10]
- O. Perron (1918, zur Zeit im Felde): *Ein neuer Existenzbeweis für die Integrale eines Systems gewöhnlicher Differentialgleichungen*. Math. Annalen, Vol.78, p.378-384. [I.7]
- O. Perron (1930): *Ueber ein vermeintliches Stabilitätskriterium*. Nachrichten Göttingen, (1930) p.28-29 (see also Fort.d.Math. 1930 I, p.380.) [I.13]
- H.J. Pesch, see H.J. Oberle & H.J. Pesch, S. Miesbach & H.J. Pesch.
- D. Pfenniger (1990): *Stability of the Lagrangian points in stellar bars*. Astron. Astrophys., Vol.230, p.55-66. [II.16]
- D. Pfenniger, see also T. de Zeeuw & D. Pfenniger.
- E. Picard (1890): *Mémoire sur la théorie des équations aux dérivées partielles et la méthode des approximations successives*. J. de Math. pures et appl., 4e série, Vol.6, p.145-210. [I.8]
- E. Picard (1891-96): *Traité d'Analyse*. 3 vols. Paris. [I.7], [I.8]
- R. Piessens, E. de Doncker-Kapenga, C.W. Überhuber & D.K. Kahaner (1983): *QUADPACK. A subroutine package for automatic integration*. Springer Series in Comput. Math., Vol.1, 301pp. [II.10]
- P. Piotrowsky (1969): *Stability, consistency and convergence of variable k -step methods for numerical integration of large systems of ordinary differential equations*. Lecture Notes in Math., 109, Dundee 1969, p.221-227. [III.5]
- H. Poincaré (1881,82,85): *Mémoire sur les courbes définies par les équations différentielles*. J. de Math., 3e série, t.7, p.375-422, 3e série, t.8, p.251-296, 4e série, t.1, p.167-244, Oeuvres t.1, p.3-84, 90-161. [I.12], [I.16]
- H. Poincaré (1892,1893,1899): *Les méthodes nouvelles de la mécanique céleste*. Tome I 385pp., Tome II 480pp., Tome III 414pp., Gauthier-Villars Paris. [I.6], [I.16], [I.14], [II.8]
- S.D. Poisson (1835): *Théorie mathématique de la chaleur*. Paris, Bachelier, 532pp., Supplément 1837, 72pp. [I.15]
- B. Van der Pol (1926): *On "Relaxation Oscillations"*. Phil. Mag., Vol.2, p.978-992; reproduced in: B. van der Pol, Selected Scientific Papers, Vol.I, North. Holland Publ. Comp. Amsterdam (1960). [I.16]
- G. Pólya & G. Szegő (1925): *Aufgaben und Lehrsätze aus der Analysis*. Two volumes, Springer Verlag; many later editions and translations. [II.9]
- P. Pouzet (1963): *Etude en vue de leur traitement numérique des équations intégrales de type Volterra*. Rev. Français Traitement Information (Chiffres), Vol.6, p.79-112. [II.17]
- P.J. Prince & J.R. Dormand (1981): *High order embedded Runge-Kutta formulae*. J. Comp. Appl. Math., Vol.7, p.67-75. [II.5]
- P.J. Prince, see also J.R. Dormand, M.E.A. El-Mikkawy & P.J. Prince, J.R. Dormand & P.J. Prince, R.W. Brankin, I. Gladwell, J.R. Dormand, P.J. Prince & W.L. Seward.
- H. Prüfer (1926): *Neue Herleitung der Sturm-Liouvillschen Reihenentwicklung stetiger Funktionen*. Math. Annalen, Vol.95, p.499-518. [I.15]

- D.I. Pullin & P.G. Saffman (1991): *Long-time symplectic integration: the example of four-vortex motion*. Proc. R. Soc. London, A, Vol.432, p.481-494. [II.16]
- Qin Meng-Zhao & Zhu Wen-Jie (1991): *Canonical Runge-Kutta-Nyström (RKN) methods for second order ordinary differential equations*. Computers Math. Applic., Vol.22, p.85-95. [II.16]
- Qin Meng-Zhao & Zhu Wen-Jie (1992): *Construction of higher order symplectic schemes by composition*. Computing, Vol.47, p.309-321. [II.16]
- Qin Meng-zhao, see also Feng Kang, Wu Hua-mo, Qin Meng-zhao & Wang Dao-liu.
- R. Radau (1880): *Étude sur les formes d'approximation qui servent à calculer la valeur numérique d'une intégrale définie*. Liouville J. de Mathém. pures et appl., 3eser., tome VI, p.283-336. (Voir p.307). [II.7]
- A. Ralston (1962): *Runge-Kutta methods with minimum error bounds*. Math. Comput., Vol.16, p.431-437, corr., Vol.17, p.488. [II.1], [II.3], [III.7]
- L. Rández, see M. Calvo, J.I. Montijano & L. Rández.
- Lord Rayleigh (1883): *On maintained vibrations*. Phil. Mag. Ser.5, Vol.15, p.229-235. [I.16]
- L.J. Reed, see R. Pearl & L.J. Reed.
- W.T. Reid (1980): *Sturmian theory for ordinary differential equations*. Springer Verlag, Appl. Math., Serie31, 559pp. [I.15]
- C. Reinsch, see J.H. Wilkinson & C. Reinsch.
- R. Reissig (1954): *Erzwungene Schwingungen mit zäher und trockener Reibung*. Math. Nachrichten, Vol.11, p.345-384; see also p.231. [II.6]
- P. Rentrop (1985): *Partitioned Runge-Kutta methods with stiffness detection and stepsize control*. Numer. Math., Vol.47, p.545-564. II.15
- J. Riccati (1712): *Soluzione generale del Problema inverso intorno à raggi osculatori..., determinar la curva, a cui convenga una tal'espressione*. Giornale de' Letterati d'Italia, Vol.11, p.204-220. [I.3]
- J. Riccati (1723): *Animadversiones in aequationes differentiales secundi gradus*. Acta Erud. Lips., anno MDCCXXIII, p.502-510. [I.3]
- L.F. Richardson (1910): *The approximate arithmetical solution by finite differences of physical problems including differential equations, with an application to the stresses in a masonry dam*. Phil. Trans., A, Vol.210, p.307-357. [II.4]
- L.F. Richardson (1927): *The deferred approach to the limit*. Phil. Trans., A, Vol.226, p.299-349. [II.4], [II.9]
- R.D. Richtmyer, see P.D. Lax & R.D. Richtmyer.
- B. Riemann (1854): *Ueber die Darstellbarkeit einer Function durch eine trigonometrische Reihe*. Von dem Verfasser behufs seiner Habilitation an der Universität zu Göttingen der philosophischen Facultät eingereicht; collected works p. 227-265. [I.6]
- W. Romberg (1955): *Vereinfachte numerische Integration*. Norske Vid. Selsk. Forhdl, Vol.28, p.30-36. [II.8], [II.9]
- E. Rothe (1930): *Zweidimensionale parabolische Randwertaufgaben als Grenzfall eindimensionaler Randwertaufgaben*. Math. Annalen, Vol.102, p. 650-670. [I.1]

- N. Rouche, P. Habets & M. Laloy (1977): *Stability theory by Liapunov's direct method*. Appl. Math. Sci. 22, Springer Verlag, 396pp. [I.13]
- E.J. Routh (1877): *A Treatise on the stability of a given state of motions*. Being the essay to which the Adams prize was adjudged in 1877, in the University of Cambridge. London 108pp. [I.13]
- E.J. Routh (1884): *A Treatise on the dynamics of a system of rigid bodies, part I and II*. 4th edition (1st ed. 1860, 6th ed. 1897, german translation with remarks of F.Klein 1898). [I.12]
- D. Ruelle & F. Takens (1971): *On the nature of turbulence*. Commun. Math. Physics, Vol.20, p.167-192. [I.16]
- A. Ruhe, see B. Kågström & A. Ruhe.
- C. Runge (1895): *Ueber die numerische Auflösung von Differentialgleichungen*. Math. Ann., Vol.46, p.167-178. [II.1], [II.4]
- C. Runge (1905): *Ueber die numerische Auflösung totaler Differentialgleichungen*. Göttinger Nachr., p.252-257. [II.1], [II.3]
- C. Runge & H. König (1924): *Vorlesungen über numerisches Rechnen*. Grundlehren XI, Springer Verlag, 372pp. [I.8], [II.1]
- R.D. Russel, see U.M. Ascher, R.M.M. Mattheij & R.D. Russel.
- R.D. Ruth (1983): *A canonical integration technique*. IEEE Trans. Nuclear Science, Vol.NS-30, p.2669-2671. [II.16]
- H. Rutishauser (1952): *Ueber die Instabilität von Methoden zur Integration gewöhnlicher Differentialgleichungen*. ZAMP, Vol.3, p.65-74. [III.3]
- H. Rutishauser, see also F.L. Bauer, H. Rutishauser & E. Stiefel.
- P.G. Saffman, see D.I. Pullin & P.G. Saffman.
- J.M. Sanz-Serna (1988): *Runge-Kutta schemes for Hamiltonian systems*. BIT Vol.28, p.877-883. [II.16]
- J.M. Sanz-Serna (1992): *Symplectic integrators for Hamiltonian problems: an overview*. Acta Numerica, Vol.1, p.243-286. [II.16]
- J.M. Sanz-Serna (1992b): *The numerical integration of Hamiltonian systems*. In: Computational ordinary differential equations, ed. by J.R. Cash & I. Gladwell, Clarendon Press, Oxford, p.437-449. [II.16]
- J.M. Sanz-Serna & L. Abia (1991): *Order conditions for canonical Runge-Kutta schemes*. SIAM J. Numer. Anal., Vol.28, p. 1081-1096. [II.16]
- J.M. Sanz-Serna, see also M.P. Calvo & J.M. Sanz-Serna, L. Abia & J.M. Sanz-Serna.
- D. Sarafyan (1966): *Error estimation for Runge-Kutta methods through pseudo-iterative formulas*. Techn. Rep. No 14, Louisiana State Univ., New Orleans, May 1966. [II.4]
- L. Schaeffer (1884): *Zur Theorie der stetigen Funktionen einer reellen Veränderlichen*. Acta Mathematica, Vol.5, p.183-194. [I.10]
- J. Schlöder, see H.G. Bock & J. Schlöder.
- I.J. Schoenberg, see F.R. Loscalzo & I.J. Schoenberg.

- I. Schur (1909): *Ueber die charakteristischen Wurzeln einer linearen Substitution mit einer Anwendung auf die Theorie der Integralgleichungen*. Math. Ann., Vol.66, p.488-510. [I.12]
- C. Scovel, see P.J. Channell & C. Scovel.
- W.L. Seward, see R.W. Brankin, I. Gladwell, J.R. Dormand, P.J. Prince & W.L. Seward.
- L.F. Shampine (1979): *Storage reduction for Runge-Kutta codes*. ACM Trans. Math. Software, Vol.5, p.245-250. [II.5]
- L.F. Shampine (1985): *Interpolation for Runge-Kutta methods*. SIAM J. Numer. Anal., Vol.22, p.1014-1027. [II.6]
- L.F. Shampine (1986): *Some practical Runge-Kutta formulas*. Math. Comp., Vol.46, p.135-150. [II.5], [II.6]
- L.F. Shampine & L.S. Baca (1983): *Smoothing the extrapolated midpoint rule*. Numer. Math., Vol.41, p.165-175. [II.9]
- L.F. Shampine & L.S. Baca (1986): *Fixed versus variable order Runge-Kutta*. ACM Trans. Math. Softw., Vol.12, p.1-23. [II.9]
- L.F. Shampine, L.S. Baca & H.-J. Bauer (1983): *Output in extrapolation codes*. Comp. & Maths. with Appls., Vol.9, p.245-255. [II.9]
- L.F. Shampine & M.K. Gordon (1975): *Computer Solution of Ordinary Differential Equations, The Initial Value Problem*. Freeman and Company, San Francisco, 318pp. [III.7]
- L.F. Shampine & H.A. Watts (1979): *The art of writing a Runge-Kutta code. II*. Appl. Math. Comput., Vol.5, p.93-121. [II.4], [III.7]
- L.F. Shampine, H.A. Watts & S.M. Davenport (1976): *Solving nonstiff ordinary differential equations - The state of the art*. SIAM Rev., Vol.18, p.376-410. [II.6]
- L.F. Shampine, see also I. Gladwell, L.F. Shampine & R.W. Brankin, P. Bogacki & L.F. Shampine.
- E.B. Shanks (1966): *Solutions of differential equations by evaluations of functions*. Math. of Comp., Vol.20, p.21-38. [II.5]
- Shi Songling (1980): *A concrete example of the existence of four limit cycles for plane quadratic systems*. Sci. Sinica, Vol.23, p.153-158. [I.16]
- G.F. Simmons (1972): *Differential equations with applications and historical notes*. MC Graw-Hill, 465pp. [I.16]
- H.H. Simonsen (1990): *Extrapolation methods for ODE's: continuous approximations, a parallel approach*. Dr.Ing. Thesis, Norwegian Inst. Tech., Div. of Math. Sciences. [II.9]
- S.W. Skan, see R.A. Frazer, W.P. Jones & S.W. Skan.
- R. Skeel (1976): *Analysis of fixed-stepsize methods*. SIAM J. Numer. Anal., Vol.13, p.664-685. [III.4], [III.8], [III.9]
- R.D. Skeel (1979): *Equivalent forms of multistep formulas*. Math. Comput., Vol.33, p.1229-1250. [III.6]
- R.D. Skeel, see also D. Okunbor & R.D. Skeel.
- B.P. Sommeijer, see P.J. van der Houwen & B.P. Sommeijer.

- D. Sommer (1965): *Numerische Anwendung impliziter Runge-Kutta-Formeln*. ZAMM, Vol. 45, Sonderheft, p. T77-T79. [II.7]
- F. Sommer, see H. Behnke & F. Sommer.
- A. Sommerfeld (1942): *Vorlesungen über theoretische Physik*. Bd.1., Mechanik; translated from the 4th german ed.: Acad. Press. [II.10], [II.14]
- D.C. Sorensen, see R. Fletcher & D.C. Sorensen.
- J.L. Soulé, see A. Guillou & J.L. Soulé.
- M.N. Spijker (1971): *On the structure of error estimates for finite difference methods*. Numer. Math., Vol.18, pp.73-100. [III.8]
- M.N. Spijker, see also W.H. Hundsdorfer & M.N. Spijker.
- D.D. Stancu, see A.H. Stroud & D.D. Stancu.
- J.F. Steffensen (1956): *On the restricted problem of three bodies*. K. danske Vidensk. Selsk., Mat-fys. Medd. 30 Nr.18. [I.8]
- I.A. Stegun, see M. Abramowitz & I.A. Stegun.
- H.J. Stetter (1970): *Symmetric two-step algorithms for ordinary differential equations*. Computing, Vol.5, p.267-280. [II.9]
- H.J. Stetter (1971): *Local estimation of the global discretization error*. SIAM J. Numer. Anal., Vol.8, p.512-523. [II.12]
- H.J. Stetter (1973): *Analysis of discretization methods for ordinary differential equations*. Springer Verlag, Berlin-Heidelberg-New York. [II.8], [II.12], [III.2], [III.8], [III.9]
- H.J. Stetter, see also W.B. Gragg & H.J. Stetter.
- D. Stewart (1990): *A high accuracy method for solving ODEs with discontinuous right-hand side*. Numer. Math., Vol.58, p.299-328. [II.6]
- E. Stiefel, see F.L. Bauer, H. Rutishauser & E. Stiefel.
- J. Stoer, see R. Bulirsch & J. Stoer.
- C. Störmer (1907): *Sur les trajectoires des corpuscules électrisés*. Arch. sci. phys. nat., Genève, Vol.24, p.5-18, 113-158, 221-247. [III.10]
- C. Störmer (1921): *Méthodes d'intégration numérique des équations différentielles ordinaires*. C.R. congr. intern. math., Strasbourg, p.243-257. [II.14], [III.10]
- A.H. Stroud & D.D. Stancu (1965): *Quadrature formulas with multiple Gaussian nodes*. SIAM J. Numer. Anal., ser.B., Vol.2, p.129-143. [II.13]
- Ch. Sturm (1829): *Bulletin des Sciences de Férussac*. Tome XI, p.419, see also: *Algèbre de Choquet et Mayer* (1832). [I.13]
- Ch. Sturm (1836): *Sur les équations différentielles linéaires du second ordre*. Journal de Math. pures et appl. (Liouville), Vol.1, p.106-186 (see also p.253, p.269, p.373 of this volume). [I.15]
- Sun Geng (孙耿) (1992): *Construction of high order symplectic Runge-Kutta Methods*. Comput. Math., Vol.11 (1993), p.250-260. [II.16]

- Y.B. Suris (1989): *The canonicity of mappings generated by Runge-Kutta type methods when integrating the systems $\dot{x} = -\partial U/\partial x$* . Zh. Vychisl. Mat. i Mat. Fiz., vol 29, p.202-211 (in Russian); same as U.S.S.R. Comput. Maths. Phys., vol 29., p.138-144. [II.16]
- Y.B. Suris (1990): *Hamiltonian Runge-Kutta type methods and their variational formulation*. Mathematical Simulation, Vol.2, p.78-87 (Russian). [II.16]
- V. Szebehely (1967): *Theory of orbits. The restricted problem of three bodies*. Acad. Press, New York, 668pp. [II.0]
- G. Szegő, see G. Pólya & G. Szegő.
- P.G.Tait, see W. Thomson (Lord Kelvin) & P.G.Tait.
- F. Takens, see D. Ruelle & F. Takens.
- K. Taubert (1976): *Differenzenverfahren für Schwingungen mit trockener und zäher Reibung und für Regelungssysteme*. Numer. Math., Vol.26, p.379-395. [II.6]
- K. Taubert (1976): *Eine Erweiterung der Theorie von G. Dahlquist*. Computing, Vol.17, p.177-185. [III.4]
- B. Taylor (1715): *Methodus incrementorum directa et inversa*. Londini 1715. [I.6]
- W. Thomson (Lord Kelvin) & P.G.Tait (1879): *Treatise on natural philosophy (Vol.I., Part I)*. Cambridge; New edition 1890, 508pp. [I.12]
- P.G. Thomson, see W.H. Enright, K.R. Jackson, S.P. Nørsett & P.G. Thomson.
- J. Todd (1950): *Notes on modern numerical analysis, I*. Math. Tables Aids Comput., Vol.4, p.39-44. [III.3]
- W. Tollmien (1938): *Ueber die Fehlerabschätzung beim Adamsschen Verfahren zur Integration gewöhnlicher Differentialgleichungen*. ZAMM, Vol.18, p.83-90. [III.4]
- S. Tremaine, see J. Binney & S. Tremaine.
- K.W. Tu, see C.W. Gear & K.W. Tu.
- C.W. Überhuber, see R. Piessens, E. de Doncker-Kapenga, C.W. Überhuber & D.K. Kahaner.
- W. Uhlmann (1957): *Fehlerabschätzungen bei Anfangswertaufgaben gewöhnlicher Differentialgleichungssysteme I. Ordnung*. ZAMM, Vol.37, p.88-99. [I.10]
- B.T. Ulrich, see B. Giovannini, L. Weiss-Parmeggiani & B.T. Ulrich.
- R. Vaillancourt, see M. Calvé & R. Vaillancourt.
- C. Van Loan, see C. Moler & C. Van Loan.
- R.S. Varga, see G. Birkhoff & R.S. Varga,
- P.F. Verhulst (1845): *Recherches mathématiques sur la loi d'accroissement de la population*. Nuov. Mem. Acad. Roy. Bruxelles, Vol.18, p.3-38. [II.17]
- J.H. Verner (1971): *On deriving explicit Runge-Kutta methods*. Proc. Conf. on Appl. Numer. Analysis, Lecture Notes in Mathematics 228, Springer Verlag, p.340-347. [II.5]
- J.H. Verner (1978): *Explicit Runge-Kutta methods with estimates of the local truncation error*. SIAM J.Numer. Anal., Vol.15, p.772-790. [II.5]
- J.H. Verner, see also G.J. Cooper & J.H. Verner.

- L. Vietoris (1953): *Der Richtungsfehler einer durch das Adamssche Interpolationsverfahren gewonnenen Näherungslösung einer Gleichung $y' = f(x, y)$* . Oesterr. Akad. Wiss., Math.-naturw. Kl., Abt. IIa, Vol.162, p.157-167 and p.293-299. [III.4]
- R. de Vogelaere (1956): *Methods of integration which preserve the contact transformation property of the Hamiltonian equations*. Report No. 4, Dept. Mathem., Unive. of Notre Dame, Notre Dame, Ind. [II.16]
- V. Volterra (1934): *Remarques sur la Note de M. Régnier et Mlle Lambin*. C.R.Acad. Sc. t. CXCI, p.1682. See also: V.Volterra - U.d'Ancona, Les associations biologiques au point de vue mathématique, Paris 1935. [II.17]
- W. Walter (1970): *Differential and integral inequalities*. Springer Verlag 352pp., german edition 1964. [I.10]
- W. Walter (1971): *There is an elementary proof of Peano's existence theorem*. Amer. Math. Monthly, Vol.78, p.170-173. [I.7]
- Wang Dao-liu, see Feng Kang, Wu Hua-mo, Qin Meng-zhao & Wang Dao-liu.
- G. Wanner (1969): *Integration gewöhnlicher Differentialgleichungen, Lie Reihen, Runge-Kutta-Methoden*. BI Mannheim Htb. 831/831a, 182pp. [I.8]
- G. Wanner (1973): *Runge-Kutta methods with expansions in even powers of h* . Computing, Vol.11, p.81-85. [II.8]
- G. Wanner (1983): *On Shi's counter example for the 16th Hilbert problem*. Internal Rep. Sect. de Math., Univ. Genève 1982; in german in: Jahrbuch Ueberblicke Mathematik 1983, ed. Chatterji, Fenyő, Kulisch, Laugwitz, Liedl, BI Mannheim, p.9-24. [I.13], [I.16]
- G. Wanner, see also K.H. Kastlunger & G. Wanner, S.P. Nørsett & G. Wanner, E. Hairer & G. Wanner, H. Knapp & G. Wanner.
- R.F. Warming & B.J. Hyett (1974): *The modified equation approach to the stability and accuracy analysis of finite-difference methods*. J. Comp. Phys., Vol.14, p.159-179. [II.16]
- D.S. Watanabe, see C.W. Gear & D.S. Watanabe.
- H.A. Watts (1983): *Starting stepsize for an ODE solver*. J. Comp. Appl. Math., Vol.9, p.177-191. [III.4]
- H.A. Watts, see also L.F. Shampine & H.A. Watts, L.F. Shampine, H.A. Watts & S.M. Davenport.
- K. Weierstrass (1858): *Ueber ein die homogenen Functionen zweiten Grades betreffendes Theorem, nebst Anwendung desselben auf die Theorie der kleinen Schwingungen*. Monatsber. der Königl. Akad. der Wiss., 4. März 1858, Werke Bd.I, p.233-246. [I.6]
- L. Weiss-Parmeggiani, see B. Giovannini, L. Weiss-Parmeggiani & B.T. Ulrich.
- J. Weissinger (1950): *Eine verschärfte Fehlerabschätzung zum Extrapolationsverfahren von Adams*. ZAMM, Vol.30, p.356-363. [III.4]
- H. Weyl (1939): *The classical groups*. Princeton, 302pp. [I.14]
- O. Wilde (1892): *Lady Windermere's Fan, Comedy in four acts*. [I.7]
- J.H. Wilkinson (1965): *The algebraic eigenvalue problem, Monographs on numerical analysis*. Oxford, 662pp. [I.9]

- J.H. Wilkinson & C. Reinsch (1970): *Linear Algebra*. Grundlehren Band 186, Springer Verlag, 439pp. [I.12]
- J.H. Wilkinson, see also G.H. Golub & J.H. Wilkinson.
- A. Wintner & F.D. Murnaghan (1931): *A canonical form for real matrices under orthogonal transformations*. Proc. Nat. Acad. Sci. U.S.A., Vol.17, p.417-420. [I.12]
- E.M. Wright (1945): *On a sequence defined by a non-linear recurrence formula*. J. of London Math. Soc., Vol.20, p.68-73. [II.17]
- E.M. Wright (1946): *The non-linear difference-differential equation*. Quart. J. of Math., Vol.17, p.245-252. [II.17]
- E.M. Wright (1955): *A non-linear difference-differential equation*. J.f.d.r.u. angew. Math., Vol.194, p.66-87. [II.17]
- K. Wright (1970): *Some relationships between implicit Runge-Kutta collocation and Lanczos τ methods, and their stability properties*. BIT Vol.10, p.217-227. [II.7]
- H. Wronski (1810): *Premier principe des méthodes algorithmiques comme base de la technique algorithmique*. Publication refused by the Acad. de Paris (for more details see: S.Dickstein, Int. Math. Congress 1904, p.515). [I.11]
- Wu Hua-mo, see Feng Kang, Wu Hua-mo, Qin Meng-zhao & Wang Dao-liu.
- H. Yoshida (1990): *Construction of higher order symplectic integrators*. Phys. Lett. A, Vol.150, p.262-268. [II.16]
- H. Yoshida (1993): *Recent progress in the theory and application of symplectic integrators*. Celestial Mechanics Dynam. Astr., Vol.56, p.27-43. [II.16]
- T. de Zeeuw & D. Pfenniger (1988): *Potential-density pairs for galaxies*. Mon. Not. R. astr. Soc., Vol.235, p.949-995. [II.16]
- M. Zennaro (1986): *Natural continuous extensions of Runge-Kutta methods*. Math. Comput., Vol.46, p.119-133. [II.17]
- M. Zennaro, see also A. Bellen & M. Zennaro, B. Owren & M. Zennaro, Z. Jackiewicz & M. Zennaro.
- N.P Zhidkov, see I.S. Berezin & N.P Zhidkov.
- Zhu Wen-Jie, see Qin Meng-Zhao & Zhu Wen-Jie.
- J.A. Zonneveld (1963): *Automatic integration of ordinary differential equations*. Report R743, Mathematisch Centrum, Postbus 4079, 1009AB Amsterdam. Appeared in book form 1964. [II.4]
- R. Zurmühl (1948): *Runge-Kutta-Verfahren zur numerischen Integration von Differentialgleichungen n-ter Ordnung*. ZAMM, Vol.28, p.173-182. [II.14]
- R. Zurmühl (1952): *Runge-Kutta Verfahren unter Verwendung höherer Ableitungen*. ZAMM, Vol.32, p.153-154. [II.13]

Symbol Index

$A \otimes I$	tensor product 393
$\mathbf{a}(t)$	B-series coefficients 265
$B(\mathbf{a}, y)$	B-series 266
$B(p)$	quadrature order 208
$b_j(\theta)$	coefficients of continuous method 188
c_i, a_{ij}, b_j	RK coefficients 134f
C	error constant 373, 414, 471
C_{p+1}	local error constant 372, 471
$C(\eta)$	simplifying assumption 208
D	differential operator 274
$D(\xi)$	simplifying assumption 208
$D^+m(x), D_+m(x)$	Dini derivatives 56
$d_i(t)$	difference set 266
$dp_i \wedge dq_i$	exterior product 101
E	principal part of S 437
$e(t)$	error coefficients 158
$e_p(x)$	global error coefficient 216f, 451f
fe	number of function evaluations 140, 252, 299, 428
$F(t)(y)$	elementary differential 146, 148, 287, 292, 305
$g_j(n)$	variable step size coefficients 398
$H(p, q)$	Hamilton function 32, 100, 312
$K_q(s)$	Peano-kernel 375
$l = (l_0, l_1, \dots)$	Nordsieck coefficients 411, 417
L	Lipschitz constant 36, 37, 54, 391
\mathcal{L}	Lagrange function 30
$L(y, x, h)$	difference operator 369, 467
LNT_q	labelled N-trees 287
LS_q	special labelled trees 150
LT_q, LT	labelled trees 146, 266
LTP_q^a, LTP^a	labelled P-trees 304, 306
NT_q	N-trees 288
p	order of the method 134, 284, 437

$P(\mathbf{c}, y)$	P-series 307
$P(EC)^M$	predictor-corrector 360, 433
$P(EC)^M E$	predictor-corrector 360, 432
$\ Q\ $	matrix norm 53
$R(x, x_0)$	resolvent 65
s	number of stages 134
S	matrix of general linear method 431
$s_i(t)$	subtree 266
t_{21}, t_{31}, \dots	trees 148
$[t_1, \dots, t_m]$	composite tree 152
$_a[t_1, \dots, t_m]$	composite tree 304
T_q, T	rooted trees 147, 266
$T_{i,1}, T_{j,k}$	extrapolation tableau 224, 225, 295
TP_q^a, TP^a	P-trees 304, 306
$V(y_1, \dots, y_n)$	Liapunov function 86
$W(x)$	Wronskian 65
$\ y\ $	vector norm 52
$y_h(x)$	numerical solution 36, 216, 391
$y(x, x_0, y_0)$	solution 95
$z_n = z(x_n, h)$	correct value function 431
α_j, β_j	multistep coefficients 368, 467
$\alpha(t)$	coefficient 147, 148, 288, 292, 304
$\beta_j(n)$	variable step size coefficients 399
$\gamma(t)$	order products 148, 151, 152, 292
$\delta^j f[x_n, \dots, x_{n-j}]$	divided differences 397
$\nabla^j f_n$	backward differences 357
$\mu(Q)$	logarithmic norm 61
$\varphi(h)$	starting procedure 431
$\Phi_j(t)$	weights 151, 289, 308
$\Phi_j(n)$	divided differences 399
$\Phi_j^*(n)$	divided differences 398
$\Phi(x_0, y_0, h)$	increment function 159, 216, 393, 431
$\Phi^*(x, y, h)$	adjoint method 219, 457
$\varrho(t)$	order of a tree 146f, 292, 304
$\varrho(\zeta), \sigma(\zeta)$	generating polynomials 370, 467
τ, τ_a	one-vertex trees (not one-tree hill!) 147, 304
ω^2	differential 2-form 102, 312

Subject Index

- Adams methods 357, 359
 - as Nordsieck methods 410, 417
 - error constant 373
 - variable step size 397
- Adjoint matrix 70
- Adjoint method 219
 - asymptotic expansion 220
 - general linear methods 457
- Aitken-Neville algorithm 226
- AREN 245
- ARES 296
- Arenstorf orbits 130, 186, 245
- Asymptotic expansion 216
 - general linear methods 448f
 - in h^2 222, 230, 459
 - second order equations 471
- Asymptotic solutions
 - for small parameters 114
- Asymptotic stability 87
- Autonomous systems 69

- B-series 266
- Backward analysis 333
- Backward differences 357
- Backward differentiation formulas (see BDF)
- BDF 364f
 - as Nordsieck methods 417, 419
 - stability 380f
 - variable step size 400, 405, 409
- Bernoulli equation 15
- Bernoulli numbers 414
- Bessel equation 24
- Boot-strapping process 191
- Boundary conditions 105
- Boundary value problems 105
- Brachystochrone problem 7, 15, 23
- BRUS 248
- Brusselator 115, 170, 248
 - full 117
 - with diffusion 248
- Bulirsch sequence 226
- Butcher barriers 173
- Butcher's Lobatto formulas 210
- Butcher's methods of order 2s 208f

- Campbell-Baker-Hausdorff formula 334
- Canonical equations 33, 100, 312
- Characteristic equation 17, 70
 - delay equations 343
- Characteristic polynomial 70
- Chemical reactions 115
- Clairaut differential equation 9f
- Collocation methods 211
 - equivalence to RK 212
 - order 212
 - second order equations 301
 - symplectic 315
 - with multiple nodes 275
- Collocation polynomial 211
- Composition of B-series 264
- Composition of RK methods 264
- Consistency conditions 371
 - general linear methods 437
 - second order equations 468
- Constant coefficients 69
 - geometric representation 76
 - numerical computations 72
- Continuous RK methods 188
 - for delay equations 342
 - for DOPRI5 191
- Convergence
 - Euler's method 35
 - general linear methods 438
 - multistep methods 391, 395
 - Nyström methods 285
 - RK methods 156
 - second order equations 468

- variable step size multistep 407
- Convergence monitor 234
- Correct value function 431
- Corrector 360
- Cowell and Crommelin's method 462
- CPEN 297
- Critical points 77
- Cyclic multistep methods 433
- D-stability 380
- Dahlquist barrier (first) 378, 384
- DEABM 423
- Defect 57
- Delay differential equations 339
 - stability 343
- Dense output 188
 - for extrapolation methods 237, 295
 - parallel 261
 - for RK methods 188
- Derivatives
 - with respect to initial values 95
 - with respect to parameters 93
 - numerical 200
- Diagonal implicit RK-method 205
- Diagonalization 69
- DIFEX1 233
- Difference equation 28
- Differential equation of Laplace 141, 356
- Differential forms 101
- Differential inequalities 56
 - for systems 63
- DIFSUB 426
- Dini derivatives 56
- DIRK-method 205
- Discontinuous equations
 - (numerical study) 196
- Discrete Laplace transformation 413
- Divided differences 397
- DOPRI5 178, 253, 477
- DOP853 181, 185, 253, 254, 481
- Dormand and Prince methods 178, 181
 - continuous extension 191, 194
 - second order equations 294
- Drops 142
- DVERK 253
- Effective order 270
- Ehle's methods 215
- Eigenfunctions 109
- Eigenvalue 69
- Eigenvector 69

- Elementary differential 146, 148, 287, 292, 305
- Embedded RK formulas 165
 - of high order 180
- Encke's method 462
- End-vertex 287
- Enzyme kinetics 348
- EPISODE 426
- Equivalence of
 - RK methods 273
 - labelled trees 147
 - N-trees 287
 - P-trees 303, 304
- ERK 134, 205
- Error
 - global 159, 216
 - local 160, 368, 436
- Error coefficients 158
- Error constant 372, 373
 - Nordsieck methods 414
 - second order equations 471
- Error estimate 58, 156
 - of Euler's method 40
 - practical 164
- Estimation of the global error 159
- Euler polygons 36
 - convergence 37
 - error estimate 40
 - for systems 52
- Euler's method 35, 51, 358
 - implicit (backward) 204, 360
- Euler-Maclaurin formula 218
- EULR 244
- Event location 195
- Existence theorem 35
 - for systems of equations 51
 - of Peano 41
 - using iteration methods 44
 - using Taylor series 46
- Explicit Adams methods 357f
- Explicit RK methods 134
 - arbitrary order 232
 - high order 173
- Exterior product 101
- External differentiation 200
- External stages 434
- Extrapolation methods 224, 261
 - as Runge-Kutta methods 232, 241
 - order 225
 - second order equations 294
 - with symmetric methods 226

- Faà di Bruno's formula 149
- Father 146
- Fehlberg's methods 177, 180
 - multiderivative 278, 309
- Feigenbaum cascades 124
 - number 125
- First integrals 34, 318
- Flow 97
- Forward step procedure 431
- Fourier 29
- FSAL (first same as last) 167
- Fundamental lemma 58
- Galactic dynamics 319
- Gauss methods 208f, 315
- GBS method 228, 294
- General linear methods 430, 434
 - convergence 438
 - order 437
 - order conditions 441
 - stability 436
- Generating functions for the γ_i 358
- Generating polynomials
 - of multistep methods 370, 467
- Gerschgorin's theorem 89
- Gill's method 139
- Global error 159, 216
 - estimation 270
- Gragg's method 228
- Greek-Roman transformation 384
- Gronwall lemma 62
- Hammer and Hollingsworth's method 205, 207
- Hamilton function (Hamiltonian) 32
 - preservation 34, 318, 325, 333
 - for galactic problem 320
 - separable 327
- Hamilton mechanics 32
- Hanging string 27
- Harmonic oscillator 312
- Harmonic sequence 226
- Heat conduction 29, 107
- Hermite interpolation 190, 261, 275
- Heun's method 135
- Higher derivative methods 274
- Hilbert's 16th problem 126
- Hopf bifurcation 117
- Hybrid methods 430
- Hypergeometric functions 22
- Immunology 349
- Implementation of multistep methods 421
- Implicit Adams methods 359f
 - as Nordsieck methods 410, 417
- Implicit differential equation 8
- Implicit midpoint rule 204, 205
- Implicit output 195
- Implicit RK methods 204, 205
 - as collocation methods 212
 - based on Gaussian formulas 208
 - based on Lobatto quadrature 210
 - existence of solution 206
- Increment function 159
 - general linear methods 431
- Index equation 21
- Infectious disease modelling 347
- Infinite series 4
- Inhomogeneous linear equation 12
 - systems 66
- Initial value 2
- Integro-differential equations 351
- Internal differentiation 201
- Internal stages 434
- Interpolation error, Control 240
- Inverse tangent problem 6
- IRK-method 205
- Irreducible methods 374
- Jacobian elliptic functions 240
- Jordan canonical form 73
- Josephson junctions 118, 119
- Kronecker tensor product 393
- Kuntzmann's methods of order 2s 208, 209
- Kutta's 3/8 rule 138
- Labelled N-tree 286
- Labelled P-tree 304
- Labelled tree 146
- Lady Windermere's Fan 39, 96, 160, 395
- Lagrange function 30
 - for galactic problem 320
- Lagrangian mechanics 30
- Large parameters 113
- Liapunov functions 86
- Limit cycle 111, 117
 - existence proof 112
 - unicity 127
- Linear differential equations 16

- homogeneous 16
- inhomogeneous 12, 17, 18, 66
- systems 64
- weak singularities 20
- with constant coefficients 16, 69
- Linear multistep formulas 356, 368
 - convergence of order p 395
 - general 368
- Lipschitz condition 37, 391
 - one-sided 60
- Local error 156, 216
 - general linear methods 436
 - multistep 368
 - numerical estimation 422
- Local extrapolation 178
- Logarithmic norm 61, 63
- Lorenz model 120, 123, 245
- LRNZ 245
- LSODE 426
- Madam Imhof's cheese pie 383
- Majorant method 47
- Matrix norms 53
- Mechanics
 - Lagrangian 30
 - Hamiltonian 32
- Merson's method 167
- Method of lines 3, 248
- Midpoint rule 132, 363
 - implicit 204
- Milne-Simpson methods 363
- Multi-step multi-stage multi-derivative methods 435
- Multiderivative methods 274, 277, 435
 - order conditions 280
 - symplectic 338
- Multiple characteristic values 17, 19
- Multistep formula
 - as general linear method 431
 - characteristic equation 378
 - cyclic 433
 - generalized 430
 - irreducible 374
 - modified 430
 - parasitic solution 379
 - Peano kernel 375
 - stability 378, 380
 - symmetric 387
 - variable step size 397, 401
- N-tree of order q 288
- Newton's interpolation formula 357, 397
- Nonlinear variation-of-constants formula 96
- Nordsieck methods 410f
 - as general linear methods 433
 - equivalence with multistep 412
- Nordsieck vector 410, 418
- Norm 52
 - logarithmic 61
 - of a matrix 53
- Normal matrices 79
- Numerical examples
 - comparisons of codes 244
 - extrapolation methods 227, 231
 - 4th order methods 140, 171
 - multistep codes 421f
 - second order equations 296
- Numerov's method 464
- Nyström methods
 - construction 291
 - convergence 285
 - general linear methods 435
 - multistep 362
 - order conditions 290, 309
 - symplectic 330
- Obreschkoff methods 277
- ODEX 236, 254, 482
- ODEX2 296, 298, 484
- One-sided Lipschitz condition 60
- One-step methods 129, 216
- Optimal formulas 139
- Optimal order
 - extrapolation 233
 - multistep 423
- Order
 - Adams methods 371, 372
 - extrapolation methods 225
 - general linear methods 437
 - labelled tree 146
 - multistep methods 370
 - RK methods 134
 - variable step size multistep 401
- Order barriers
 - RK methods (Butcher) 173, 179
 - parallel RK methods 259
 - multistep (Dahlquist) 384, 468
- Order conditions
 - general linear methods 441
 - multiderivative methods 280
 - multistep 368, 370

- number of 154
- Nyström methods 290, 292
- RK methods 143, 145, 153, 269
- Order control
 - extrapolation 233, 236
 - multistep 423
- Oregonator 118, 119
- Orthogonal matrix 70
- P-optimal methods 259
- P-series 307
- P-trees 304, 306
- Parallel methods 257
- Parasitic solution 379
- Partial differential equations 3, 248
- Partitioned RK method 302
 - symplectic 326
- Partitioned systems 279, 302
- Peano kernel 375
- Pendulum equation 126, 335
- Perimeter of the ellipse 24
- Periodic solution 111
- Phase-space 77
- Picard-Lindelöf iteration 45
 - for systems 54
- PLEI 245
- Pleiades problem 245
- Poincaré sections 112, 121
 - computations 195
- Population dynamics 345
- Preconsistency conditions 441
- Predictor-corrector process 360, 432
- Principal error term 158
- Principle of the argument 382
- Propagation of sound 27
- Pseudo Runge-Kutta methods 430
- q-derivative RK method 274
- Quasimonotone 63
- Radau scheme 204
- Rational extrapolation 227
- Recurrence relations for the γ_i 358, 359
- Regular singular point 22
- Reliability 261
- Resolvent 65
- RETARD 343, 488
- Retarded arguments 339
- Riccati equation 44
- Richardson extrapolation 164
- Rigorous error bounds 156
- RKF45 251
- RKN6 & RKN12 298
- Romberg sequence 225
- Root of a tree 146
- Root condition 380
- ROPE 247
- Rothe's method 3
- Rounding errors 242, 472
- Routh tableau 84
 - computation 85
- Routh-Hurwitz criterion 81
- Runge's methods 134
- Runge-Kutta methods 132
 - delay 341
 - diagonal implicit 205
 - explicit 134f
 - general formulation 134, 143
 - implicit 205, 205
 - of order four 135
 - singly diagonal implicit 205
 - symplectic 315
 - "the" 138
 - two-step method 446
 - violating (1.9) 308
- Schur decomposition 70
- Schur-Cohn criterion 388
- SDIRK-method 205
 - order three 207
- Second order equations 13,
 - collocation 301
 - extrapolation methods 294
 - multistep methods 461, 467f
 - Nyström methods 283f
 - Runge-Kutta methods 283
- Separable Hamiltonian 327
- Shooting method 107
- Simplifying assumptions 136, 208, 444
- Singularities 20
- SN-trees 291, 292
- Son 146
- Son-father mapping 146
- Special labelled trees 150
- Spherical pendulum 31, 33, 103
- Stability
 - asymptotic 87
 - BDF 380
 - delay equations 343
 - general linear methods 436
 - multistep formula 378
 - non-autonomous systems 88

- nonlinear systems 87
- second order equations 467
- variable step size multistep 402
- Stable in the sense of Liapunov 80
- Starting procedure 356, 431
- Starting step size 169
- Steady-state approximations 113
- Steam engine governor 91
- Step size control 167
 - extrapolation methods 233, 236
 - multistep methods 421
 - numerical study 170, 236
- Step size freeze 203
- Step size ratio 402, 421
- Stetter's error estimation 270
- Stoermer's methods 295, 462, 464
- Strange attractors 120
- Strictly stable methods 450
- Sturm sequence 81, 82
- Sturm's comparison theorem 107, 108
- Sturm-Liouville eigenvalue problems 107
- Subordinate matrix norms 53
- Switching function 196
- Symmetric methods 221
 - asymptotic expansion 222
 - general linear methods 459
 - multistep 387
- Symplectic
 - mapping 102, 104
 - multiderivative RK 338
 - Nyström method 330
 - partitioned method 326
 - RK methods 312, 316
- Systems of equations 26
 - autonomous 69
 - linear 64
 - second order 283
 - with constant coefficients 69
- Taylor expansion
 - of exact solutions 144, 148
 - of RK solutions 133, 144, 151
- Taylor series 46
 - convergence proof 46
 - recursive computation 47
- Three body problem 129
- Three-eighth's rule 138
- Time lags 339
- Total differential equation 12
- Trapezoidal rule 204
- Tree 147
 - number of 147
- Two-step RK method 446
- Unitary matrix 70
- Van der Pol's equation 111
- Variable step size
 - Adams 397
 - BDF 400
 - multistep methods 401
- Variation of constants 18, 66
 - nonlinear 96
- Variational Calculus 7
- Variational equation 95
- Vector notation 3, 52
- Vibrating string 27
- VODE 426
- Volume-preserving flows 97
- Weak singularities 20
 - for systems 68
 - RK methods applied to 163
- Weakly stable methods 454
- Work-precision diagrams 140, 171, 252f, 299f, 428f, 466
- WPLT 298
- Wronskian 19, 65
- Zero-stability 380
- Zonneveld's method 167

Springer Series in Computational Mathematics

14

Editorial Board

R.L. Graham, La Jolla (CA)

J. Stoer, Würzburg

R. Varga, Kent (Ohio)

Springer-Verlag
Berlin Heidelberg GmbH

E. Hairer G. Wanner

Solving Ordinary Differential Equations II

Stiff and Differential-Algebraic Problems

Second Revised Edition

With 137 Figures



Springer

Ernst Hairer
Gerhard Wanner
Université de Genève
Section de Mathématiques, C.P. 240
2-4 rue du Lièvre
CH-1211 Genève 24
Switzerland
e-mail: Ernst.Hairer|Gerhard.Wanner@math.unige.ch

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Solving ordinary differential equations / E. Hairer; G. Wanner.

Bd. 1 verf. von E. Hairer, S.P. Norsett und G. Wanner
NE: Hairer, Ernst; Norsett, Syvert P.; Wanner, Gerhard
2. Stiff and differential algebraic problems. - 2., rev. ed. - 1996
(Springer Series in Computational mathematics; 14)
ISBN 978-3-642-05220-0 ISBN 978-3-642-05221-7 (eBook)
DOI 10.1007/978-3-642-05221-7

NE: GT

Corrected Second Printing 2002

The cover design of the Springer Series in Computational Mathematics is based on Figure 10.4 on page 149 of this book.

Mathematics Subject Classification (2000): 65Lxx, 34A50

ISBN 978-3-642-05220-0

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag Berlin Heidelberg GmbH.

Violations are liable for prosecution under the German Copyright Law.

springeronline.com

© Springer-Verlag Berlin Heidelberg 1991, 1996

Originally published by Springer-Verlag Berlin Heidelberg New York in 1996

Softcover reprint of the hardcover 2nd edition 1996

The use of general descriptive names, registered names, trademarks etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Cover design: *design&production*, Heidelberg

Camera-ready copy by the authors

Printed on acid-free paper

46/3142ck-54321

To Evi and Myriam

From the Preface to the First Edition

“Whatever regrets may be, we have done our best.”
(Sir Ernest Shackleton, turning back on 9 January 1909 at 88° 23' South.)

Brahms struggled for 20 years to write his first symphony. Compared to this, the 10 years we have been working on these two volumes may even appear short.

This second volume treats stiff differential equations and differential algebraic equations. It contains three chapters: Chapter IV on one-step (Runge-Kutta) methods for stiff problems, Chapter V on multistep methods for stiff problems, and Chapter VI on singular perturbation and differential-algebraic equations.

Each chapter is divided into sections. Usually the first sections of a chapter are of an introductory nature, explain numerical phenomena and exhibit numerical results. Investigations of a more theoretical nature are presented in the later sections of each chapter.

As in Volume I, the formulas, theorems, tables and figures are numbered consecutively in each section and indicate, in addition, the section number. In cross references to other chapters the (latin) chapter number is put first. References to the bibliography are again by “author” plus “year” in parentheses. The bibliography again contains only those papers which are discussed in the text and is in no way meant to be complete.

It is a pleasure to thank J. Butcher, G. Dahlquist, and S.P. Nørsett (coauthor of Volume I) for their interest in the subject and for the numerous discussions we had with them which greatly inspired our work. Special thanks go to the participants of our seminar in Geneva, in particular Ch. Lubich, A. Ostermann and M. Roche, where all the subjects of this book have been presented and discussed over the years. Much help in preparing the manuscript was given by J. Steinig, Ch. Lubich and A. Ostermann who read and re-read the whole text and made innumerable corrections and suggestions for improvement. We express our sincere gratitude to them. Many people have seen particular sections and made invaluable suggestions and remarks: M. Crouzeix, P. Deuffhard, K. Gustafsson, G. Hall, W. Hundsdorfer, L. Jay, R. Jeltsch, J.P. Kauthen, H. Kraaijevanger, R. März, and O. Nevanlinna. . . . Several pictures were produced by our children Klaudia Wanner and Martin Hairer, the one by drawing the other by hacking.

The marvellous, perfect and never failing TEX program of D. Knuth allowed us to deliver a camera-ready manuscript to Springer Verlag, so that the book could be produced rapidly and at a reasonable price. We acknowledge with pleasure the numerous remarks of the planning and production group of Springer Verlag concerning fonts, style and other questions of elegance.

Preface to the Second Edition

The preparation of the second edition allowed us to improve the first edition by rewriting many sections and by eliminating errors and misprints which have been discovered. In particular we have included new material on

- methods with extended stability (Chebyshev methods) (Sect. IV.2);
- improved computer codes and new numerical tests for one- and multistep methods (Sects. IV.10 and V.5);
- new results on properties of error growth functions (Sects. IV.11 and IV.12);
- quasilinear differential equations with state-dependent mass matrix (Sect. VI.6).

We have completely reorganized the chapter on differential-algebraic equations by including three new sections on

- index reduction methods (Sect. VII.2);
- half-explicit methods for index-2 systems (Sect. VII.6);
- symplectic methods for constrained Hamiltonian systems and backward error analysis on manifolds (Sect. VII.8).

Our sincere thanks go to many persons who have helped us with our work:

- all readers who kindly drew our attention to several errors and misprints in the first edition, in particular C. Bendtsen, R. Chan, P. Chartier, T. Eirola, L. Jay, P. Kaps, J.-P. Kauthen, P. Leone, S. Maset, B. Owren, and L.F. Shampine;
- those who read preliminary versions of the new parts of this edition for their invaluable suggestions: M. Arnold, J. Cash, D.J. Higham, P. Kunkel, Chr. Lubich, A. Medovikov, A. Murua, A. Ostermann, and J. Verwer.
- the staff of the Geneva computing center and of the mathematics library for their constant help;
- the planning and production group of Springer-Verlag for numerous suggestions on presentation and style.

All figures have been recomputed and printed, together with the text, in Postscript. All computations and text processings were done on the SUN workstations of the Mathematics Department of the University of Geneva.

April 1996

The Authors

Contents

Chapter IV. Stiff Problems – One-Step Methods

IV.1 Examples of Stiff Equations	2
Chemical Reaction Systems	3
Electrical Circuits	4
Diffusion	6
A “Stiff” Beam	8
High Oscillations	11
Exercises	11
IV.2 Stability Analysis for Explicit RK Methods	15
Stability Analysis for Euler’s Method	15
Explicit Runge-Kutta Methods	16
Extrapolation Methods	18
Analysis of the Examples of IV.1	18
Automatic Stiffness Detection	21
Step-Control Stability	24
A PI Step Size Control	28
Stabilized Explicit Runge-Kutta Methods	31
Exercises	37
IV.3 Stability Function of Implicit RK-Methods	40
The Stability Function	40
A-Stability	42
L-Stability and $A(\alpha)$ -Stability	44
Numerical Results	46
Stability Functions of Order $\geq s$	47
Padé Approximations to the Exponential Function	48
Exercises	49
IV.4 Order Stars	51
Introduction	51
Order and Stability for Rational Approximations	56
Stability of Padé Approximations	58
Comparing Stability Domains	58
Rational Approximations with Real Poles	61
The Real-Pole Sandwich	62
Multiple Real-Pole Approximations	67
Exercises	70
IV.5 Construction of Implicit Runge-Kutta Methods	71
Gauss Methods	71
Radau IA and Radau IIA Methods	72

Lobatto IIIA, IIIB and IIIC Methods	75
The W -Transformation	77
Construction of Implicit Runge-Kutta Methods	83
Stability Function	84
Positive Functions	86
Exercises	89
IV.6 Diagonally Implicit RK Methods	91
Order Conditions	91
Stiffly Accurate SDIRK Methods	92
The Stability Function	96
Multiple Real-Pole Approximations with $R(\infty)=0$	98
Choice of Method	99
Exercises	100
IV.7 Rosenbrock-Type Methods	102
Derivation of the Method	102
Order Conditions	104
The Stability Function	108
Construction of Methods of Order 4	108
Higher Order Methods	111
Implementation of Rosenbrock-Type Methods	111
The “Hump”	113
Methods with Inexact Jacobian (W -Methods)	114
Exercises	117
IV.8 Implementation of Implicit Runge-Kutta Methods	118
Reformulation of the Nonlinear System	118
Simplified Newton Iterations	119
The Linear System	121
Step Size Selection	123
Implicit Differential Equations	127
An SDIRK-Code	128
SIRK-Methods	128
Exercises	130
IV.9 Extrapolation Methods	131
Extrapolation of Symmetric Methods	131
Smoothing	133
The Linearly Implicit Mid-Point Rule	134
Implicit and Linearly Implicit Euler Method	138
Implementation	139
Exercises	142
IV.10 Numerical Experiments	143
The Codes Used	143
Twelve Test Problems	144
Results and Discussion	152
Partitioning and Projection Methods	160
Exercises	165
IV.11 Contractivity for Linear Problems	167
Euclidean Norms (Theorem of von Neumann)	168
Error Growth Function for Linear Problems	169
Small Nonlinear Perturbations	172
Contractivity in $\ \cdot\ _\infty$ and $\ \cdot\ _1$	175
Study of the Threshold Factor	176

Absolutely Monotonic Functions	178
Exercises	179
IV.12 B-Stability and Contractivity	180
One-Sided Lipschitz Condition	180
B -Stability and Algebraic Stability	181
Some Algebraically Stable IRK Methods	183
AN -Stability	184
Reducible Runge-Kutta Methods	187
The Equivalence Theorem for S -Irreducible Methods	188
Error Growth Function	193
Computation of $\varphi_B(x)$	195
Exercises	199
IV.13 Positive Quadrature Formulas and B-Stable RK-Methods ..	201
Quadrature Formulas and Related Continued Fractions	201
Number of Positive Weights	203
Characterization of Positive Quadrature Formulas	205
Necessary Conditions for Algebraic Stability	206
Characterization of Algebraically Stable Methods	209
The “Equivalence” of A - and B -Stability	211
Exercises	213
IV.14 Existence and Uniqueness of IRK Solutions	215
Existence	215
A Counterexample	217
Influence of Perturbations and Uniqueness	218
Computation of $\alpha_0(A^{-1})$	220
Methods with Singular A	222
Lobatto IIIC Methods	223
Exercises	223
IV.15 B-Convergence	225
The Order Reduction Phenomenon	225
The Local Error	228
Error Propagation	229
B -Convergence for Variable Step Sizes	230
B -Convergence Implies Algebraic Stability	232
The Trapezoidal Rule	234
Order Reduction for Rosenbrock Methods	236
Exercises	237

Chapter V. Multistep Methods for Stiff Problems

V.1 Stability of Multistep Methods	240
The Stability Region	240
Adams Methods	242
Predictor-Corrector Schemes	244
Nyström Methods	245
BDF	246
The Second Dahlquist Barrier	247
Exercises	249
V.2 “Nearly” A-Stable Multistep Methods	250
$A(\alpha)$ -Stability and Stiff Stability	250
High Order $A(\alpha)$ -Stable Methods	251
Approximating Low Order Methods with High Order Ones	253

A Disc Theorem	254
Accuracy Barriers for Linear Multistep Methods	254
Exercises	259
V.3 Generalized Multistep Methods	261
Second Derivative Multistep Methods of Enright	261
Second Derivative BDF Methods	265
Blended Multistep Methods	266
Extended Multistep Methods of Cash	267
Multistep Collocation Methods	270
Methods of "Radau" Type	273
Exercises	275
V.4 Order Stars on Riemann Surfaces	279
Riemann Surfaces	279
Poles Representing Numerical Work	283
Order and Order Stars	284
The "Daniel and Moore Conjecture"	286
Methods with Property C'	288
General Linear Methods	290
Dual Order Stars	295
Exercises	297
V.5 Experiments with Multistep Codes	300
The Codes Used	300
Exercises	304
V.6 One-Leg Methods and G-Stability	305
One-Leg (Multistep) Methods	305
Existence and Uniqueness	306
G -Stability	307
An Algebraic Criterion	309
The Equivalence of A -Stability and G -Stability	310
A Criterion for Positive Functions	313
Error Bounds for One-Leg Methods	314
Convergence of A -Stable Multistep Methods	317
Exercises	319
V.7 Convergence for Linear Problems	321
Difference Equations for the Global Error	321
The Kreiss Matrix Theorem	323
Some Applications of the Kreiss Matrix Theorem	326
Global Error for Prothero and Robinson Problem	328
Convergence for Linear Systems with Constant Coefficients	329
Matrix Valued Theorem of von Neumann	330
Discrete Variation of Constants Formula	332
Exercises	337
V.8 Convergence for Nonlinear Problems	339
Problems Satisfying a One-Sided Lipschitz Condition	339
Multiplier Technique	342
Multipliers and Nonlinearities	346
Discrete Variation of Constants and Perturbations	348
Convergence for Nonlinear Parabolic Problems	349
Exercises	354
V.9 Algebraic Stability of General Linear Methods	356
G -Stability	356

Algebraic Stability	357
AN -Stability and Equivalence Results	359
Multistep Runge-Kutta Methods	362
Simplifying Assumptions	363
Quadrature Formulas	365
Algebraically Stable Methods of Order $2s$	366
B -Convergence	368
Exercises	370

Chapter VI. Singular Perturbation Problems and Index 1 Problems

VI.1 Solving Index 1 Problems	372
Asymptotic Solution of van der Pol's Equation	372
The ε -Embedding Method for Problems of Index 1	374
State Space Form Method	375
A Transistor Amplifier	376
Problems of the Form $Mu' = \varphi(u)$	378
Convergence of Runge-Kutta Methods	380
Exercises	381
VI.2 Multistep Methods	382
Methods for Index 1 Problems	382
Convergence for Singular Perturbation Problems	383
Exercises	387
VI.3 Epsilon Expansions for Exact and RK Solutions	388
Expansion of the Smooth Solution	388
Expansions with Boundary Layer Terms	389
Estimation of the Remainder	391
Expansion of the Runge-Kutta Solution	392
Convergence of RK-Methods for Differential-Algebraic Systems	394
Existence and Uniqueness of the Runge-Kutta Solution	397
Influence of Perturbations	398
Estimation of the Remainder in the Numerical Solution	399
Numerical Confirmation	403
Perturbed Initial Values	405
Exercises	406
VI.4 Rosenbrock Methods	407
Definition of the Method	407
Derivatives of the Exact Solution	408
Trees and Elementary Differentials	409
Taylor Expansion of the Exact Solution	411
Taylor Expansion of the Numerical Solution	412
Order Conditions	415
Convergence	416
Stiffly Accurate Rosenbrock Methods	418
Construction of RODAS, a Stiffly Accurate Embedded Method	420
Inconsistent Initial Values	422
Exercises	424
VI.5 Extrapolation Methods	426
Linearly Implicit Euler Discretization	426
Perturbed Asymptotic Expansion	428
Order Tableau	431

Error Expansion for Singular Perturbation Problems	433
Dense Output	438
Exercises	441
VI.6 Quasilinear Problems	442
Example: Moving Finite Elements	442
Problems of Index One	445
Numerical Treatment of $C(y)y' = f(y)$	446
Extrapolation Methods	447
Exercises	448

Chapter VII. Differential-Algebraic Equations of Higher Index

VII.1 The Index and Various Examples	452
Linear Equations with Constant Coefficients	452
Differentiation Index	454
Differential Equations on Manifolds	457
The Perturbation Index	459
Control Problems	461
Mechanical Systems	463
Exercises	465
VII.2 Index Reduction Methods	468
Index Reduction by Differentiation	468
Stabilization by Projection	470
Differential Equations with Invariants	472
Methods Based on Local State Space Forms	474
Overdetermined Differential-Algebraic Equations	477
Unstructured Higher Index Problems	478
Exercises	480
VII.3 Multistep Methods for Index 2 DAE	481
Existence and Uniqueness of Numerical Solution	482
Influence of Perturbations	484
The Local Error	485
Convergence for BDF	486
General Multistep Methods	489
Solution of the Nonlinear System by Simplified Newton	490
Exercises	491
VII.4 Runge-Kutta Methods for Index 2 DAE	492
The Nonlinear System	492
Estimation of the Local Error	494
Convergence for the y -Component	496
Convergence for the z -Component	497
Collocation Methods	498
Superconvergence of Collocation Methods	500
Projected Runge-Kutta Methods	502
Summary of Convergence Results	504
Exercises	505
VII.5 Order Conditions for Index 2 DAE	506
Derivatives of the Exact Solution	506
Trees and Elementary Differentials	507
Taylor Expansion of the Exact Solution	508

Derivatives of the Numerical Solution	510
Order Conditions	512
Simplifying Assumptions	514
Projected Runge-Kutta Methods	515
Exercises	518
VII.6 Half-Explicit Methods for Index 2 Systems	519
Half-Explicit Runge-Kutta Methods	520
Extrapolation Methods	525
β -Blocked Multistep Methods	527
Exercises	529
VII.7 Computation of Multibody Mechanisms	530
Description of the Model	530
Fortran Subroutines	533
Computation of Consistent Initial Values	535
Numerical Computations	536
A Stiff Mechanical System	541
Exercises	542
VII.8 Symplectic Methods for Constrained Hamiltonian Systems ..	543
Properties of the Exact Flow	544
First Order Symplectic Method	545
SHAKE and RATTLE	548
The Lobatto IIIA-IIIB Pair	550
Composition Methods	554
Backward Error Analysis (for ODEs)	555
Backward Error Analysis on Manifolds	559
Exercises	562
Appendix. Fortran Codes	565
Driver for the Code RADAU5	566
Subroutine RADAU5	568
Subroutine RADAUP	574
Subroutine RODAS	574
Subroutine SEULEX	575
Problems with Special Structure	575
Use of SOLOUT and of Dense Output	576
Bibliography	577
Symbol Index	605
Subject Index	607

Chapter IV. Stiff Problems – One-Step Methods

This chapter introduces stiff (styv (Swedish first!), steif (German), stíf (Islandic), stijf (Dutch), raide (French), rígido (Spanish), rígido (Portuguese), stiff (Italian), kankea (Finnish), δύσκαμπτο (Greek), merev (Hungarian), rigid (Rumanian), tog (Slovenian), čvrst (Serbo-Croatian), tuhý (Czecho-Slovakian), sztywny (Polish), jäik (Estonian), stiegrs (Latvian), standus (Lithuanian), stign (Breton), zurrun (Basque), sert (Turkish), жесткий (Russian), ТВЪРД (Bulgarian), קשיח (Hebrew), ساق (Arabic), بَنَظَنَت (Urdu), سخت (Persian), कठिण (Sanskrit), कठु (Hindi), 剛性 (Chinese), 硬い (Japanese), cở cứng (Vietnamese), ngumu (Swaheli) . . .) differential equations. While the intuitive meaning of stiff is clear to all specialists, much controversy is going on about it's correct mathematical definition (see e.g. p.360-363 of Aiken (1985)). The most pragmatistical opinion is also historically the first one (Curtiss & Hirschfelder 1952): *stiff equations are equations where certain implicit methods, in particular BDF, perform better, usually tremendously better, than explicit ones*. The eigenvalues of the Jacobian $\partial f / \partial y$ play certainly a role in this decision, but quantities such as the dimension of the system, the smoothness of the solution or the integration interval are also important (Sections IV.1 and IV.2).

Stiff equations need new concepts of stability (A-stability, Sect. IV.3) and lead to mathematical theories on order restrictions (order stars, Sect. IV.4). Stiff equations require implicit methods; we therefore focus in Sections IV.5 and IV.6 on implicit Runge-Kutta methods, in IV.7 on (semi-implicit) Rosenbrock methods and in IV.9 on semi-implicit extrapolation methods. The actual efficient implementation of implicit Runge-Kutta methods poses a number of problems which are discussed in Sect. IV.8. Section IV.10 then reports on some numerical experience for all these methods.

With Sections IV.11, IV.12 and IV.13 we begin with the discussion of contractivity (B -stability) for linear and nonlinear differential equations. The chapter ends with questions of existence and numerical stability of the implicit Runge-Kutta solutions (Sect. IV.14) and a convergence theory which is independent of the stiffness (B -convergence, Sect. IV.15).

IV.1 Examples of Stiff Equations

... Around 1960, things became completely different and everyone became aware that the world was full of stiff problems.
(G. Dahlquist in Aiken 1985)

Stiff equations are problems for which explicit methods don't work. Curtiss & Hirschfelder (1952) explain stiffness on one-dimensional examples such as

$$y' = -50(y - \cos x). \quad (1.1)$$

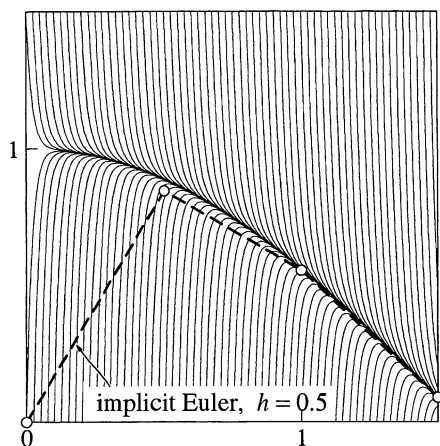


Fig. 1.1. Solution curves of (1.1) with implicit Euler solution

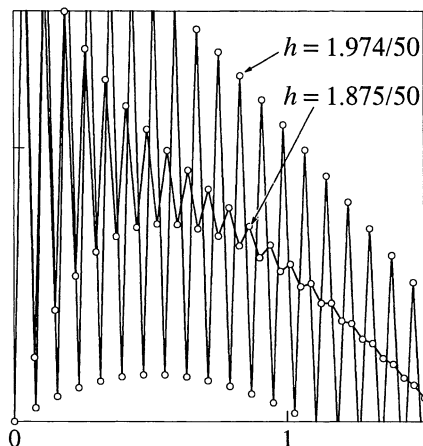


Fig. 1.2. Explicit Euler for $y(0) = 0$, $h = 1.974/50$ and $1.875/50$

Solution curves of Equation (1.1) are shown in Fig. 1.1. There is apparently a smooth solution in the vicinity of $y \approx \cos x$ and all other solutions reach this one after a rapid “transient phase”. Such transients are typical of stiff equations, but are neither sufficient nor necessary. For example, the solution with initial value $y(0) = 1$ (more precisely $2500/2501$) has *no* transient. Fig. 1.2 shows Euler polygons for the initial value $y(0) = 0$ and step sizes $h = 1.974/50$ (38 steps) and $h = 1.875/50$ (40 steps). We observe that whenever the step size is a little too large (larger than $2/50$), the numerical solution goes too far beyond the equilibrium and violent oscillations occur.

Looking for better methods for differential equations such as (1.1), Curtiss and Hirschfelder discovered the BDF method (see Sect. III.1): the approximation

$y \approx \cos x$ (i.e., $f(x, y) = 0$) is only a crude approximation to the smooth solution, since the derivative of $\cos x$ is not zero. It is much better, for a given solution value y_n , to search for a point y_{n+1} where the slope of the vector field is directed towards y_n , hence

$$\frac{y_{n+1} - y_n}{h} = f(x_{n+1}, y_{n+1}). \quad (1.2)$$

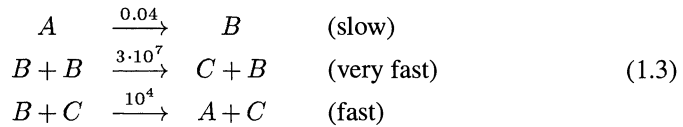
This is the implicit Euler method. The dotted line in Fig. 1.1 consists of three implicit Euler steps and demonstrates impressively the good stability property of this method. Equation (1.1) is thus apparently “stiff” in the sense of Curtiss and Hirschfelder.

Extending the above idea “by taking higher order polynomials to fit y at a large number of points” then leads to the BDF methods.

Chemical Reaction Systems

When the equations represent the behaviour of a system containing a number of fast and slow reactions, a forward integration of these equations becomes difficult. (H.H. Robertson 1966)

The following example of Robertson’s (1966) has become very popular in numerical studies (Willoughby 1974):



which leads to the equations

$$\begin{array}{lll} \text{A:} & y_1' = -0.04y_1 + 10^4 y_2 y_3 & y_1(0) = 1 \\ \text{B:} & y_2' = 0.04y_1 - 10^4 y_2 y_3 - 3 \cdot 10^7 y_2^2 & y_2(0) = 0 \\ \text{C:} & y_3' = 3 \cdot 10^7 y_2^2 & y_3(0) = 0. \end{array} \quad (1.4)$$

After a bad experience with explicit Euler just before, let’s try a higher order method and a more elaborate code for this example: DOPRI5 (cf. Volume 1). The numerical solutions obtained for y_2 with $Rtol = 10^{-2}$ (209 steps) as well as with $Rtol = 10^{-3}$ (205 steps) and $Atol = 10^{-6} \cdot Rtol$ are displayed in Fig. 1.3. Fig. 1.4 presents the step sizes used by the code and also the local error estimates. There, all rejected steps are crossed out.

We observe that the solution y_2 rapidly reaches a quasi-stationary position in the vicinity of $y_2' = 0$, which in the beginning ($y_1 = 1, y_3 = 0$) is at $0.04 \approx 3 \cdot 10^7 y_2^2$, hence $y_2 \approx 3.65 \cdot 10^{-5}$, and then very slowly goes back to zero again. The numerical method, however, integrates this smooth solution by thousands of apparently unnecessary steps. Moreover, the chosen step sizes are more or less independent of the chosen tolerance. Hence, they seem to be governed by stability

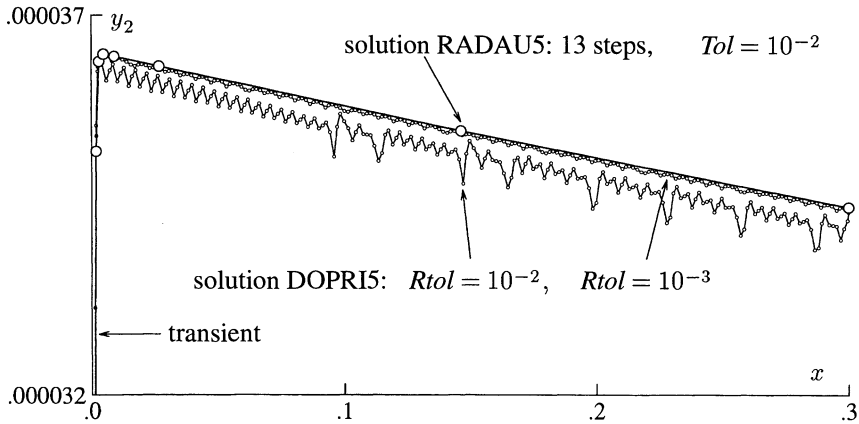


Fig. 1.3. Numerical solution for problem (1.4) with DOPRI5 and RADAU5

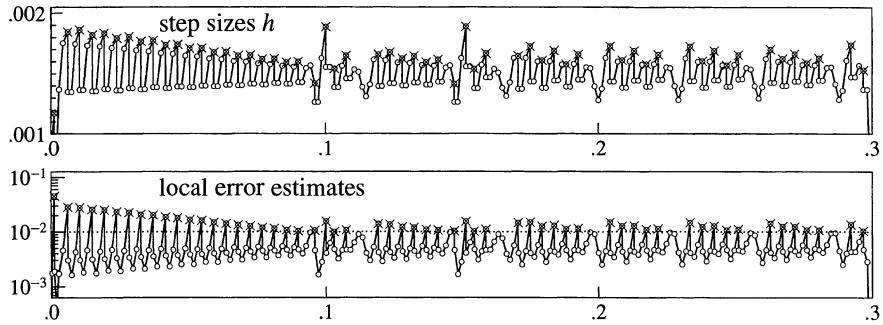


Fig. 1.4. Step sizes and local error estimates of DOPRI5, $Tol = 10^{-2}$

rather than by precision requirements. It can also be seen that an implicit Runge-Kutta code (such as RADAU5 described in Sections IV.5 and IV.8) integrates this equation without any problem.

Electrical Circuits

This behavior is known, at least in part, to any experienced worker in the field.
(G. Hall 1985)

One of the simplest nonlinear equations describing a circuit is van der Pol's equation (see Sect. I.16)

$$\begin{aligned} y_1' &= y_2 & y_1(0) &= 2 \\ y_2' &= \mu(1 - y_1^2)y_2 - y_1 & y_2(0) &= 0. \end{aligned} \quad (1.5)$$

We have seen in Chapter II that this equation is easily integrated for moderate values of μ . But we now choose $\mu = 500$ and suspect that the problem might

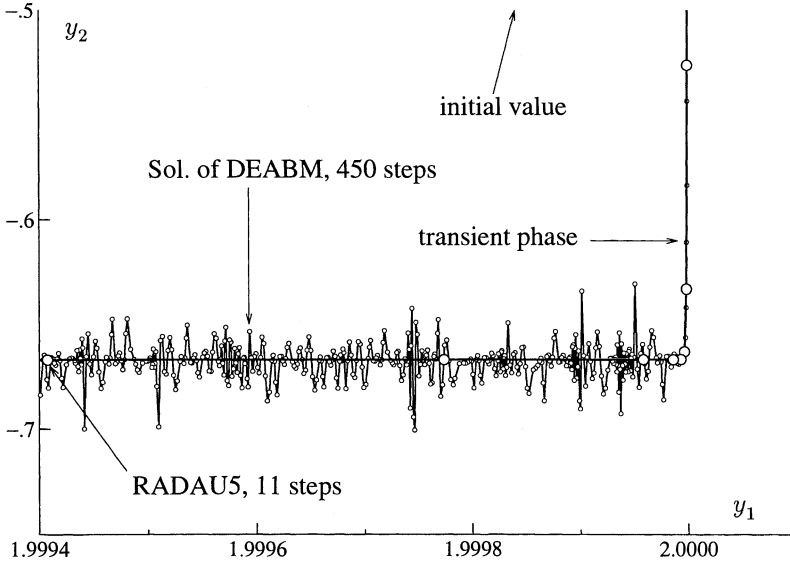


Fig. 1.5. Numerical solution for DEABM at equation (1.5'), $Rtol = 10^{-2}$, $Atol = 10^{-7}$

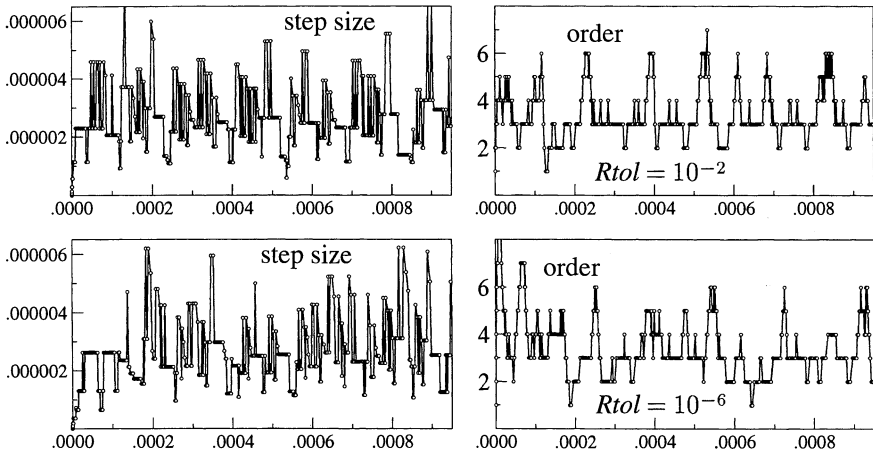


Fig. 1.6. Step sizes and orders for DEABM, $Rtol = 10^{-2}$, 10^{-6} , $Atol = 5 \cdot 10^{-8}$

become difficult. It turns out that the period of the solution increases with μ . We therefore rescale the solutions and introduce $t = x/\mu$, $z_1(t) = y_1(x)$, $z_2(t) = \mu y_2(x)$. In the resulting equation the factor μ^2 multiplies the entire second line of f . Substituting again y for z , x for t and $\mu^2 = 1/\varepsilon$ we obtain

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= \mu^2((1 - y_1^2)y_2 - y_1) \end{aligned} \quad \text{or} \quad \begin{aligned} y_1' &= y_2 \\ \varepsilon y_2' &= (1 - y_1^2)y_2 - y_1. \end{aligned} \quad (1.5')$$

The steady-state approximation (see Vol. I, Formula (I.16.5)) then becomes independent of μ .

Why not try a multistep code this time? For example the predictor-corrector Adams code DEABM of Shampine & Watts. Figures 1.5 and 1.6 show the numerical solution, the step sizes and the orders for the first 450 steps. Eventually the code stops with the message *Idid* = -4 (“the problem appears to be stiff”). The implicit Runge-Kutta code RADAU5 integrates over the same interval in 11 steps.

Diffusion

Stalling numerical processes must be wrong.
(A “golden rule” of Achi Brandt)

Another source of stiffness is the translation of diffusion terms by divided differences (method of lines, see Sect. I.1) into a large system of ODE's. We choose the Brusselator (see (16.12) of Sect. I.16) in one spatial variable x

$$\begin{aligned}\frac{\partial u}{\partial t} &= A + u^2 v - (B + 1)u + \alpha \frac{\partial^2 u}{\partial x^2} \\ \frac{\partial v}{\partial t} &= Bu - u^2 v + \alpha \frac{\partial^2 v}{\partial x^2}\end{aligned}\tag{1.6}$$

with $0 \leq x \leq 1$, $A = 1$, $B = 3$, $\alpha = 1/50$ and boundary conditions

$$\begin{aligned}u(0, t) &= u(1, t) = 1, & v(0, t) &= v(1, t) = 3, \\ u(x, 0) &= 1 + \sin(2\pi x), & v(x, 0) &= 3.\end{aligned}$$

We replace the second spatial derivatives by finite differences on a grid of N points $x_i = i/(N + 1)$ ($1 \leq i \leq N$), $\Delta x = 1/(N + 1)$ and obtain from (1.6)

$$\begin{aligned}u'_i &= 1 + u_i^2 v_i - 4u_i + \frac{\alpha}{(\Delta x)^2} (u_{i-1} - 2u_i + u_{i+1}), \\ v'_i &= 3u_i - u_i^2 v_i + \frac{\alpha}{(\Delta x)^2} (v_{i-1} - 2v_i + v_{i+1}), \\ u_0(t) &= u_{N+1}(t) = 1, & v_0(t) &= v_{N+1}(t) = 3, \\ u_i(0) &= 1 + \sin(2\pi x_i), & v_i(0) &= 3, \quad i = 1, \dots, N.\end{aligned}\tag{1.6'}$$

Table 1.1. Results for (1.6') with ODEX for $0 \leq t \leq 10$

N	Tol	accepted steps	rejected steps	function calls
10	10^{-4}	21	3	365
20	10^{-4}	81	25	1138
30	10^{-4}	167	45	2459
40	10^{-4}	275	62	4316
40	10^{-2}	266	59	3810

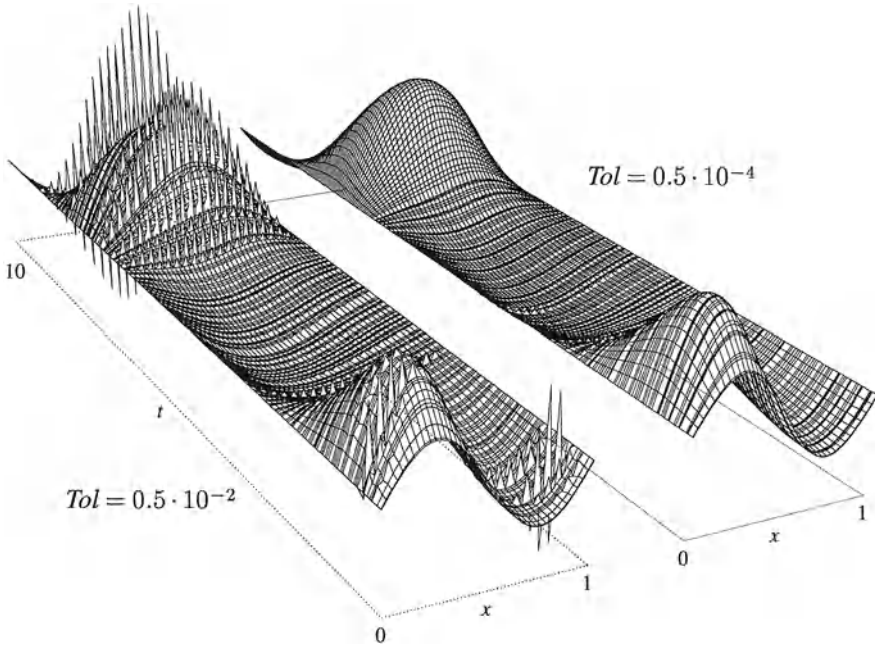


Fig. 1.7. Solution $u(x, t)$ of (1.6') with $N = 40$ using ODEX

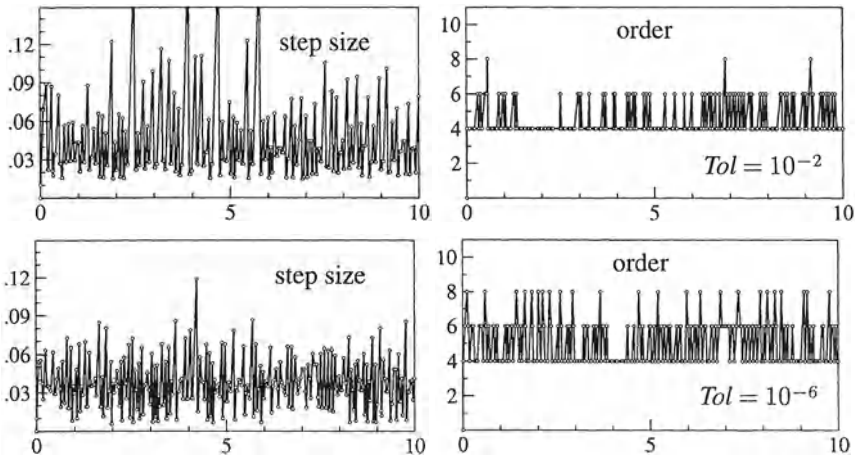


Fig. 1.8. Step size and order of ODEX at (1.6') with $N = 40$

This time we try the extrapolation code ODEX (see Volume I) and integrate over $0 \leq t \leq 10$ with $Atol = Rtol = Tol$. The number of necessary steps increases curiously with N , as is shown in Table 1.1. Again, for N large, the computing time is nearly independent of the desired tolerance, the computed solutions, however, differ considerably (see Fig. 1.7). Even the smooth 10^{-4} -solution shows curious

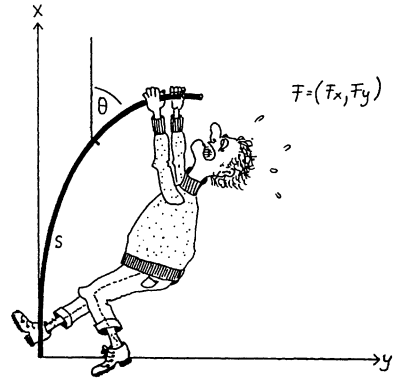
stripes which are evidently unconnected with the behaviour of the solution. Fig. 1.8 shows the extremely ragged step size and order changes which take place in this example.

We again have all the characteristics of a “stiff” problem, and the use of an implicit method promises better results. However, when applying such a method, one must carefully take advantage of the banded or sparse structure of the Jacobian matrix. Otherwise the numerical work involved in the linear algebra would increase with N^3 , precisely as the work for the explicit method (N^2 for the number of steps and N for the work per step).

A “Stiff” Beam

Although it is common to talk about “stiff differential equations,” an equation per se is not stiff, a particular initial value problem for that equation may be stiff, in some regions, but the sizes of these regions depend on the initial values and the error tolerance.
(C.W. Gear 1982)

Let us conclude our series of examples by a problem from mechanics: the motion of an elastic beam. We suppose the beam inextensible of length 1 and thin. So we neglect shearing forces and rotatory inertia. We further want to allow it arbitrarily large movements. Thus, the most natural coordinate system to use is the angle θ as a function of arc length s and time t . We further suppose the beam clamped at $s = 0$ and a force $\vec{F} = (F_x, F_y)$ acting at the free end $s = 1$. The beam is then described by the equations



(Drawing by K. Wanner)

$$x(s, t) = \int_0^s \cos \theta(\sigma, t) d\sigma, \quad y(s, t) = \int_0^s \sin \theta(\sigma, t) d\sigma. \quad (1.7)$$

In order to obtain the equations of motion for this problem, we apply Lagrange theory (Lagrange 1788). This requires that we form $L = T - U$ where T is the kinetic and U the potential energy. For the first of these we have simply

$$T = \frac{1}{2} \int_0^1 ((\dot{x}(s, t))^2 + (\dot{y}(s, t))^2) ds. \quad (1.8)$$

The potential energy is made up of energy from bending (depending on the curvature) and from exterior forces as follows:

$$U = \frac{1}{2} \int_0^1 (\theta'(s, t))^2 ds - F_x(t)x(1, t) - F_y(t)y(1, t). \quad (1.9)$$

Here dots and primes denote derivatives with respect to t and s respectively. The equations of motion are now obtained by a “trivial” calculation (we are grateful to our colleague J. Descloux for having shown us how this must be done!) using the Hamilton principle which leads to (see Exercise 2)

$$\begin{aligned} & \int_0^1 G(s, \sigma) \cos(\theta(s, t) - \theta(\sigma, t)) \ddot{\theta}(\sigma, t) d\sigma \\ &= \theta''(s, t) + \cos \theta(s, t) F_y(t) - \sin \theta(s, t) F_x(t) \\ & - \int_0^1 G(s, \sigma) \sin(\theta(s, t) - \theta(\sigma, t)) (\dot{\theta}(\sigma, t))^2 d\sigma, \quad 0 \leq s \leq 1 \end{aligned} \quad (1.10)$$

$$\theta(0, t) = 0, \quad \theta'(1, t) = 0 \quad (1.11)$$

where

$$G(s, \sigma) = 1 - \max(s, \sigma) \quad (1.12)$$

is Green's function for the problem $-w''(s) = g(s)$, $w'(0) = w(1) = 0$. If we discretize the integrals with the help of the midpoint rule

$$\int_0^1 f(\theta(\sigma, t)) d\sigma = \frac{1}{n} \sum_{k=1}^n f(\theta_k), \quad \theta_k = \theta\left(\left(k - \frac{1}{2}\right) \frac{1}{n}, t\right), \quad k = 1, \dots, n \quad (1.13)$$

Equations (1.10) become

$$\begin{aligned} \sum_{k=1}^n a_{lk} \ddot{\theta}_k &= n^4 (\theta_{l-1} - 2\theta_l + \theta_{l+1}) + n^2 (\cos \theta_l F_y - \sin \theta_l F_x) \\ & - \sum_{k=1}^n g_{lk} \sin(\theta_l - \theta_k) \dot{\theta}_k^2, \quad l = 1, \dots, n \end{aligned} \quad (1.10')$$

$$\theta_0 = -\theta_1, \quad \theta_{n+1} = \theta_n \quad (1.11')$$

where

$$a_{lk} = g_{lk} \cos(\theta_l - \theta_k), \quad g_{lk} = n + \frac{1}{2} - \max(l, k). \quad (1.14)$$

Integration without preparation is frustration.
(Reverend Leon Sullivan)

Numerical integration of (1.10') seems quite tedious, since the acceleration $\ddot{\theta}$ is only given implicitly. The computation of $\ddot{\theta}_k$ requires the solution of a linear

system $A\ddot{\theta} = v$. Due to the special structure of A , this can be done efficiently, since with $B = (b_{lk})$, $b_{lk} = g_{lk} \sin(\theta_l - \theta_k)$, we have

$$A + iB = \text{diag}(e^{i\theta_1}, \dots, e^{i\theta_n}) G \text{diag}(e^{-i\theta_1}, \dots, e^{-i\theta_n}). \quad (1.15)$$

The matrix $G = (g_{lk})$ has the beautiful inverse

$$G^{-1} = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 3 \end{pmatrix}, \quad (1.16)$$

a positive definite tridiagonal matrix (a natural coincidence: G^{-1} represents the second order difference operator, and G comes from the Green function for a second order integration problem). Now

$$(A + iB)^{-1} = C + iD = \text{diag}(e^{i\theta_1}, \dots, e^{i\theta_n}) G^{-1} \text{diag}(e^{-i\theta_1}, \dots, e^{-i\theta_n})$$

and

$$AC - BD = I, \quad AD + BC = 0 \quad (1.17)$$

lead to $A^{-1} = C + DC^{-1}D$. We can also simplify the term $-\sum g_{lk} \sin(\theta_l - \theta_k) \dot{\theta}_k^2$, which in vector notation is $-B\dot{\theta}^2$, with the formula $A^{-1}B = -DC^{-1}$ (from (1.17)). The accelerations $\ddot{\theta}_k$ are now obtained from (1.10') as follows.

- a) Let $v_l = n^4(\theta_{l-1} - 2\theta_l + \theta_{l+1}) + n^2(\cos \theta_l F_y - \sin \theta_l F_x)$,
- b) Compute $w = Dv + \dot{\theta}^2$ (D is bidiagonal);
- c) Solve the tridiagonal system $Cu = w$,
- d) Compute $\ddot{\theta} = Cv + Du$.

Thus the evaluation of (1.10') reduces to $\mathcal{O}(n)$ operations (instead of $\mathcal{O}(n^3)$). We choose the initial conditions

$$\theta(s, 0) = 0, \quad \dot{\theta}(s, 0) = 0 \quad (1.18)$$

and apply the exterior forces

$$F_x = -\varphi(t), \quad F_y = \varphi(t), \quad \varphi(t) = \begin{cases} 1.5 \cdot \sin^2 t & 0 \leq t \leq \pi \\ 0 & \pi \leq t. \end{cases} \quad (1.19)$$

The resulting system of ODE's is then integrated for $0 \leq t \leq 5$ by the code DOP853 of Volume I, although strictly speaking, the code is of too high an order for such a problem. The results are summarized in Table 1.2.

We observe the same phenomenon as before, the number of necessary steps increases like $\mathcal{O}(n^2)$ (the numerical work like $\mathcal{O}(n^3)$), and is more or less independent of the chosen tolerance. The numerical solution for $n = 40$ is displayed in Fig. 1.9. Only each 20th of the nearly 9000 steps is drawn (otherwise the picture would just be completely black). The computed solution looks perfectly smooth and there is no apparent reason for the need of *so* many steps. In fact due to lack

Table 1.2. Results for the beam (1.10') with DOP853

n	Tol	accepted steps	rejected steps	function calls
5	10^{-7}	142	35	2091
10	10^{-7}	383	26	4884
20	10^{-7}	1397	273	19769
40	10^{-7}	6913	1347	97775
20	10^{-3}	1486	450	22784
20	10^{-5}	1967	266	26532
20	10^{-7}	1397	273	19769

of stability, the numerical method produces small vibrations which are invisible for $Tol = 10^{-7}$, and which force the integrator to such small step sizes. If we relax the high precision requirement, these oscillations become visible (Fig. 1.10).

High Oscillations

Let us now choose slightly perturbed initial values in the beam equation (1.10'). Instead of (1.18) we put

$$\theta_1 = \dots = \theta_{n-1} = 0, \quad \theta_n = 0.4, \quad \dot{\theta}_1 = \dots = \dot{\theta}_n = 0. \quad (1.18')$$

This time, the *correct* solution for $n = 10$ of (1.10') computed with $Tol = 10^{-6}$ and more than 2000 steps is displayed in Fig. 1.11.

The solution is highly oscillatory, no damping wipes out the fast vibrations since the system is conservative. Hence also an implicit method, if required to follow all these oscillations, would need the same number of steps and there would of course be no advantage in using it. So we see that the decision whether a problem should be regarded as stiff or nonstiff ("... that is the question"), may also depend on the chosen initial conditions. On the other hand, we shall see in Sect. IV.2 that whenever these high oscillations are not desired, implicit methods are a marvellous instrument for wiping them out.

Exercises

1. (Curtiss & Hirschfelder 1952). "It is interesting to notice that this method of integration (the implicit Euler) may be used in either direction". Integrate equation (1.1) *backward* with step size -0.5 and initial value $y(1.5) = 0$ in three steps. Observe that the numerical solution remains stable and follows the smooth solution.
2. Derive the equations of motion (1.10) for the elastic beam from (1.8) and (1.9).

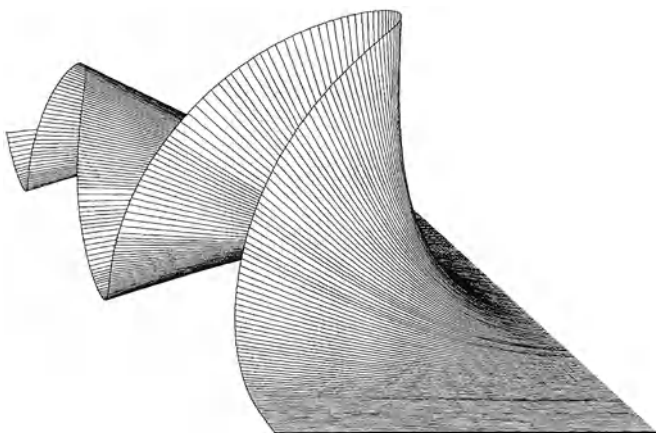


Fig. 1.9. DOP853 on the beam with $Tol = 0.0000001$, $n = 40$, every 20th step drawn.

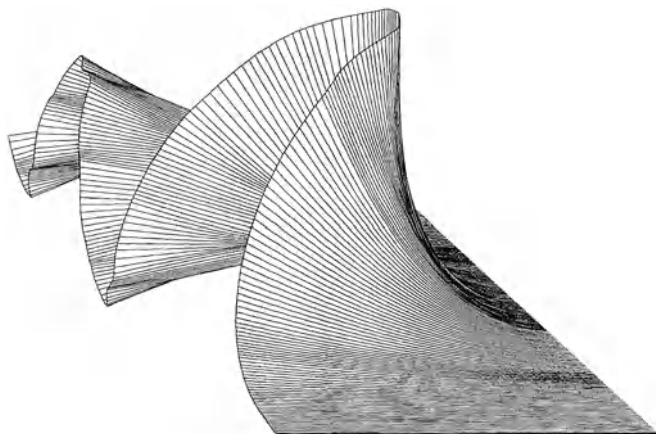


Fig. 1.10. DOP853 on the beam with $Tol = 0.0075$, $n = 20$, every 5th step drawn.

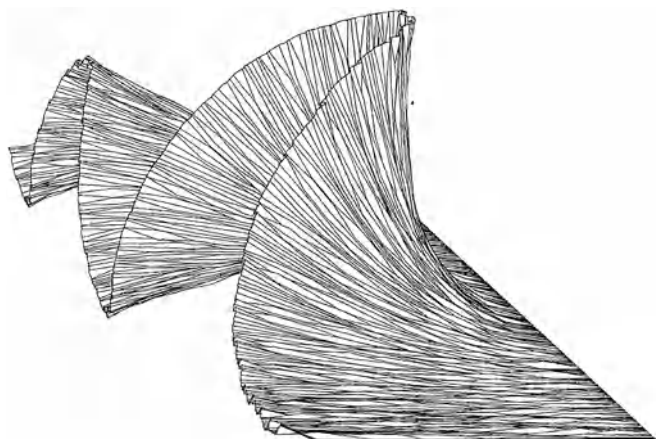


Fig. 1.11. DOP853 on highly oscillatory beam with $Tol = 0.000001$, $n = 10$, every 4th step drawn.

Hint. If you want to avoid differentiation in function spaces, then discretize the beam as, say,

$$x_j = \Delta s \sum_{k=1}^j \cos \theta_k, \quad y_j = \Delta s \sum_{k=1}^j \sin \theta_k, \quad j = 1, \dots, n, \quad \Delta s = \frac{1}{n} \quad (1.20)$$

$$T = \frac{\Delta s}{2} \sum_{j=1}^n (\dot{x}_j^2 + \dot{y}_j^2) = \frac{\Delta s}{2} \sum_{j=1}^n \dot{z}_j \dot{\bar{z}}_j, \quad z_j = \Delta s \sum_{k=1}^j e^{i\theta_k}$$

$$U = \frac{\Delta s}{2} \sum_{j=1}^n \left(\frac{\theta_j - \theta_{j-1}}{\Delta s} \right)^2 - F_x \Delta s \sum_{k=1}^n \cos \theta_k - F_y \Delta s \sum_{k=1}^n \sin \theta_k,$$

form the Lagrange function $L = T - U$ and apply n -dimensional Lagrange theory (Lagrange (1788), Vol. II, Sect. VII and VIII, a very clear derivation can be found in Sommerfeld (1942), Vol. I, §36)

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{\theta}_k} \right) - \frac{\partial L}{\partial \theta_k} = 0$$

or

$$\sum_{l=1}^n L_{\dot{\theta}_k \dot{\theta}_l} \ddot{\theta}_l = L_{\theta_k} - L_{\dot{\theta}_k t} - \sum_{l=1}^n L_{\dot{\theta}_k \theta_l} \dot{\theta}_l. \quad (1.21)$$

3. Apply an explicit code to the Oregonator (Chapter I, Equation (16.15))

$$\begin{aligned} y_1' &= 77.27 \left(y_2 + y_1 (1 - 8.375 \times 10^{-6} y_1 - y_2) \right) \\ y_2' &= \frac{1}{77.27} (y_3 - (1 + y_1) y_2) \\ y_3' &= 0.161 (y_1 - y_3) \end{aligned} \quad (1.22)$$

and study its performance.

4. a) Compute the equations of motion of the *hanging rope* (Fig. 1.12) of length 1 by using the results of Exercise 2. The potential energy has to be replaced by

$$U = - \int_0^1 x(s, t) ds.$$

Result.

$$\begin{aligned} & \int_0^1 G(s, \sigma) \cos(\theta(s, t) - \theta(\sigma, t)) \ddot{\theta}(\sigma, t) d\sigma \\ &= - \int_0^1 G(s, \sigma) \sin(\theta(s, t) - \theta(\sigma, t)) (\dot{\theta}(\sigma, t))^2 d\sigma - (1 - s) \sin \theta(s, t) \end{aligned} \quad (1.23)$$

for $0 \leq s \leq 1$, or, when discretized

$$\sum_{k=1}^n a_{lk} \ddot{\theta}_k = - \sum_{k=1}^n b_{lk} \dot{\theta}_k^2 - n \left(n + \frac{1}{2} - l \right) \sin \theta_l. \quad (1.23')$$

b) Do numerical computations with DOPRI5 or DOP853. Choose as initial position a hanging rope in equilibrium which is then released at one end.

Hint. The hanging rope in equilibrium satisfies, in the usual coordinates,

$$\int_{x_0}^{x_1} y \sqrt{1 + (y')^2} dx = \min \quad \text{with} \quad \int_{x_0}^{x_1} \sqrt{1 + (y')^2} dx = 1,$$

which, using a Lagrange multiplier, becomes

$$\int_{x_0}^{x_1} (y - \lambda) \sqrt{1 + (y')^2} dx = \text{stat}.$$

Applying (2.6) of Sect. I.2 yields $y - \lambda = K \sqrt{1 + (y')^2}$ with solution

$$y = \lambda + K \cosh \left(\frac{x + \alpha}{K} \right).$$

Suitable choices of the parameters and change of coordinates ($K = 1/2$, $\lambda = -K \cosh(\alpha/K)$, $x \rightarrow y$, $y \rightarrow -x$) then lead to

$$\theta(s, 0) = \pi/2 - \arctan(\sinh(2\alpha) - 2s). \quad (1.24)$$

Result. DOP853 has computed the solution for $0 \leq t \leq 5$, $n = 60$ and $Tol = 10^{-5}$, $\alpha = 0.6$, in 203 steps (Fig. 1.12). The number of steps increases here like $\mathcal{O}(n)$, so the rope is — evidently — less stiff than the beam.

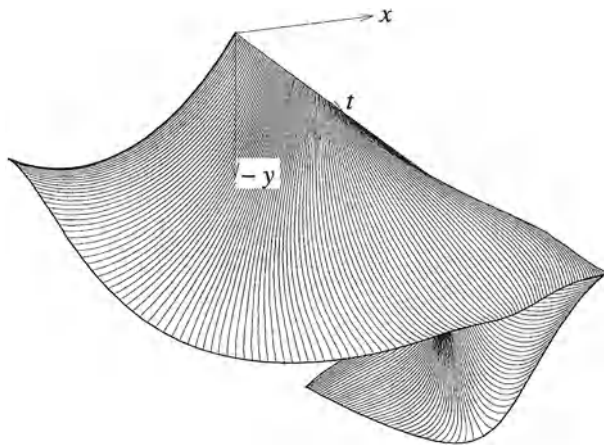


Fig. 1.12. Movement of hanging rope, every step drawn

IV.2 Stability Analysis for Explicit RK Methods

... werden wir bei dem Anfangswertproblem hyperbolischer Gleichungen erkennen, dass die Konvergenz allgemein nur dann vorhanden ist, wenn die Verhältnisse der Gittermaschen in verschiedenen Richtungen gewissen Ungleichungen genügen.

(Courant, Friedrichs & Lewy 1928)

The first analysis of instability phenomena and step size restrictions for hyperbolic equations was made in the famous paper of Courant, Friedrichs & Lewy (1928). Later, many authors undertook a stability analysis, very often independently, in order to explain the phenomena encountered in the foregoing section. An early and beautiful paper on this subject is Guillaou & Lago (1961).

Stability Analysis for Euler's Method

Let $\varphi(x)$ be a smooth solution of $y' = f(x, y)$. We linearize f in its neighbourhood as follows

$$y'(x) = f(x, \varphi(x)) + \frac{\partial f}{\partial y}(x, \varphi(x))(y(x) - \varphi(x)) + \dots \quad (2.1)$$

and introduce $y(x) - \varphi(x) = \bar{y}(x)$ to obtain

$$\bar{y}'(x) = \frac{\partial f}{\partial y}(x, \varphi(x)) \cdot \bar{y}(x) + \dots = J(x)\bar{y}(x) + \dots \quad (2.2)$$

As a first approximation we consider the Jacobian $J(x)$ as constant and neglect the error terms. Omitting the bars we arrive at

$$y' = Jy. \quad (2.2')$$

If we now apply, say, Euler's method to (2.2'), we obtain

$$y_{m+1} = R(hJ)y_m \quad (2.3)$$

with

$$R(z) = 1 + z. \quad (2.4)$$

The behaviour of (2.3) is studied by transforming J to Jordan canonical form (see Sect. I.12). We suppose that J is diagonalizable with eigenvectors v_1, \dots, v_n and write y_0 in this basis as

$$y_0 = \sum_{i=1}^n \alpha_i v_i. \quad (2.5)$$

Inserting this into (2.3) we obtain

$$y_m = \sum_{i=1}^n (R(h\lambda_i))^m \alpha_i \cdot v_i, \quad (2.6)$$

where the λ_i are the corresponding eigenvalues (see also Exercises 1 and 2). Clearly y_m remains bounded for $m \rightarrow \infty$, if for all eigenvalues the complex number $z = h\lambda_i$ lies in the set

$$S = \left\{ z \in \mathbb{C}; |R(z)| \leq 1 \right\} = \left\{ z \in \mathbb{C}; |z - (-1)| \leq 1 \right\}$$

which is the circle of radius 1 and centre -1 . This leads to the explanation of the results encountered in Example (1.1). There we have $\lambda = -50$, and $h\lambda \in S$ means that $0 \leq h \leq 2/50$, in perfect accordance with the numerical observations.

Explicit Runge-Kutta Methods

An explicit Runge-Kutta method (Sect. II.2, Formula (2.3)) applied to (2.2') gives

$$\begin{aligned} g_i &= y_m + hJ \sum_{j=1}^{i-1} a_{ij} g_j \\ y_{m+1} &= y_m + hJ \sum_{j=1}^s b_j g_j. \end{aligned} \quad (2.7)$$

Inserting g_j repeatedly from the first line, this becomes

$$y_{m+1} = R(hJ)y_m$$

where

$$R(z) = 1 + z \sum_j b_j + z^2 \sum_{j,k} b_j a_{jk} + z^3 \sum_{j,k,l} b_j a_{jk} a_{kl} + \dots \quad (2.8)$$

is a polynomial of degree $\leq s$.

Definition 2.1. The function $R(z)$ is called the *stability function* of the method. It can be interpreted as the numerical solution after one step for

$$y' = \lambda y, \quad y_0 = 1, \quad z = h\lambda, \quad (2.9)$$

the famous *Dahlquist test equation*. The set

$$S = \left\{ z \in \mathbb{C}; |R(z)| \leq 1 \right\} \quad (2.10)$$

is called the *stability domain* of the method.

Theorem 2.2. If the Runge-Kutta method is of order p , then

$$R(z) = 1 + z + \frac{z^2}{2!} + \dots + \frac{z^p}{p!} + \mathcal{O}(z^{p+1}).$$

Proof. The exact solution of (2.9) is e^z and therefore the numerical solution $y_1 = R(z)$ must satisfy

$$e^z - R(z) = \mathcal{O}(h^{p+1}) = \mathcal{O}(z^{p+1}). \quad (2.11)$$

Another argument is that the expressions in (2.8) appear in the order conditions for the “tall” trees $\tau, t_{21}, t_{32}, t_{44}, t_{59}, \dots$ (see Table 2.2 of Sect. II.2, p. 148). They are therefore equal to $1/q!$ for $q \leq p$. \square

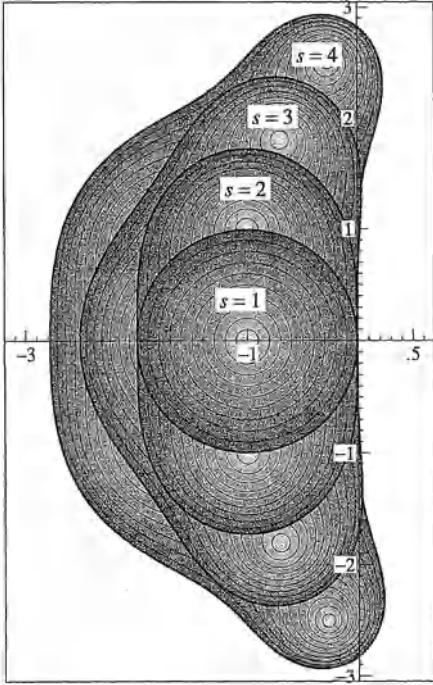


Fig. 2.1. Stability domains for explicit Runge-Kutta methods of order $p = s$

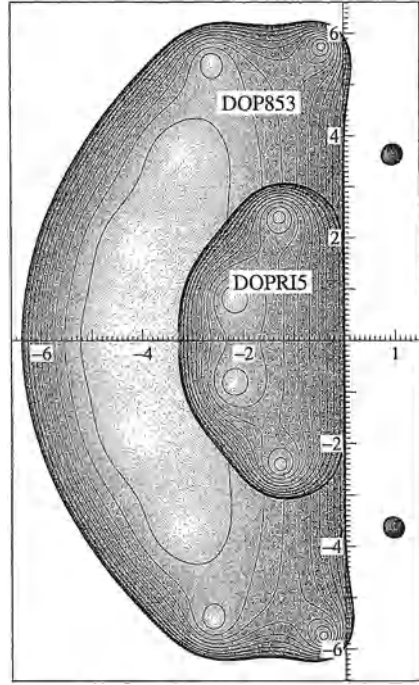


Fig. 2.2. Stability domains for DOPRI methods

As a consequence, all explicit Runge-Kutta methods with $p = s$ possess the stability function

$$R(z) = 1 + z + \frac{z^2}{2!} + \dots + \frac{z^s}{s!}. \quad (2.12)$$

The corresponding stability domains are represented in Fig. 2.1.

The method of Dormand & Prince DOPRI5 (Sect. II.5, Table 5.2) is of order 5 with $s = 6$ (the 7th stage is for error estimation only). Here $R(z)$ is obtained by direct computation. The result is

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + \frac{z^5}{120} + \frac{z^6}{600}. \quad (2.13)$$

For DOP853 (Sect. II.5, Fig. 5.3), $R(z)$ becomes

$$R(z) = \sum_{j=0}^8 \frac{z^j}{j!} + 2.6916922001691 \cdot 10^{-6} z^9 + 2.3413451082098 \cdot 10^{-7} z^{10} \\ + 1.4947364854592 \cdot 10^{-8} z^{11} + 3.6133245781282 \cdot 10^{-10} z^{12}. \quad (2.14)$$

The stability domains for these two methods are given in Fig. 2.2.

Extrapolation Methods

The GBS-algorithm (see Sect. II.9, Formulas (9.10), (9.13)) applied to $y' = \lambda y$, $y(0) = 1$ leads with $z = H\lambda$ to

$$y_0 = 1, \quad y_1 = 1 + \frac{z}{n_j} \\ y_{i+1} = y_{i-1} + 2 \frac{z}{n_j} y_i \quad i = 1, 2, \dots, n_j \\ T_{j1} = \frac{1}{4}(y_{n_j-1} + 2y_{n_j} + y_{n_j+1}) \\ T_{j,k+1} = T_{j,k} + \frac{T_{j,k} - T_{j-1,k}}{(n_j/n_{j-k})^2 - 1}. \quad (2.15)$$

The stability domains for the diagonal terms T_{22} , T_{33} , T_{44} , and T_{55} for the harmonic sequence

$$\{n_j\} = \{2, 4, 6, 8, 10, \dots\}$$

(the one which is used in ODEX) are displayed in Fig. 2.3. We have also added those for the methods *without* the smoothing step (II.9.13c), which shows some difference for negative real eigenvalues.

Analysis of the Examples of IV.1

The Jacobian for the Robertson reaction (1.3) is given by

$$\begin{pmatrix} -0.04 & 10^4 y_3 & 10^4 y_2 \\ 0.04 & -10^4 y_3 - 6 \cdot 10^7 y_2 & -10^4 y_2 \\ 0 & 6 \cdot 10^7 y_2 & 0 \end{pmatrix}$$

which in the neighbourhood of the equilibrium $y_1 = 1$, $y_2 = 0.0000365$, $y_3 = 0$ is

$$\begin{pmatrix} -0.04 & 0 & 0.365 \\ 0.04 & -2190 & -0.365 \\ 0 & 2190 & 0 \end{pmatrix}$$

with eigenvalues

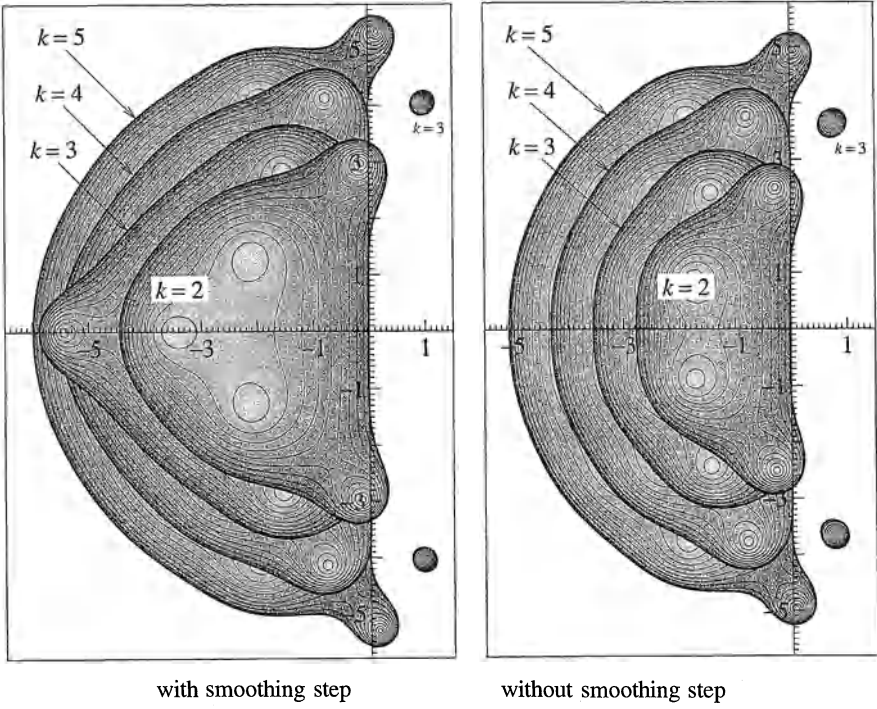


Fig. 2.3. Stability domains for GBS extrapolation methods

$$\lambda_1 = 0, \quad \lambda_2 = -0.405, \quad \lambda_3 = -2189.6.$$

The third one produces stiffness. For stability we need (see the stability domain of DOPRI5 in Fig. 2.2) $-2190h \geq -3.3$, hence $h \leq 0.0015$. This again confirms the numerical observations.

The Jacobian of example (1.6') (Brusselator reaction with diffusion) is a large $2N \times 2N$ matrix. It is composed of reaction terms and diffusion terms:

$$J = \begin{pmatrix} \text{diag}(2u_i v_i - 4) & \text{diag}(u_i^2) \\ \text{diag}(3 - 2u_i v_i) & \text{diag}(-u_i^2) \end{pmatrix} + \frac{\alpha}{(\Delta x)^2} \begin{pmatrix} K & 0 \\ 0 & K \end{pmatrix} \quad (2.16)$$

where

$$K = \begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & -2 & 1 \\ & & & 1 & -2 \end{pmatrix}. \quad (2.17)$$

The eigenvalues of K are known (see Sect. I.6, Formula (6.7b)), namely

$$\mu_k = -4 \left(\sin \frac{\pi k}{2N+2} \right)^2, \quad (2.18)$$

and therefore the double eigenvalues of the right hand matrix in (2.16) are

$$-\frac{4\alpha}{(\Delta x)^2} \left(\sin \frac{\pi k}{2N+2} \right)^2 = -4\alpha(N+1)^2 \left(\sin \frac{\pi k}{2N+2} \right)^2, \quad (2.19)$$

and are located between $-4\alpha(N+1)^2$ and 0. Since this matrix is symmetric, its eigenvalues are well conditioned and the first matrix on the right side of (2.16) with much smaller coefficients can be regarded as a small perturbation. Therefore the eigenvalues of J in (2.16) will remain close to those of the unperturbed matrix and lie in a stripe neighbouring the interval $[-4\alpha(N+1)^2, 0]$. Numerical computations for $N = 40$ show for example that the largest negative eigenvalue of J varies between -133.3 and -134.9 , while the unperturbed value is $-4 \cdot 41^2 \cdot \sin^2(40\pi/82)/50 = -134.28$. Since most stability domains for ODEX end close to -5.5 on the real axis (Fig. 2.3), this leads for $N = 40$ to $h \leq 0.04$ and the number of steps must be ≥ 250 (compare with Table 1.1).

In order to explain the behaviour of the beam equation, we linearize it in the neighbourhood of the solution $\theta_k = \dot{\theta}_k = 0$, $F_x = F_y = 0$. There (1.10') becomes

$$G\ddot{\theta} = n^4 \begin{pmatrix} -3 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & -2 & 1 \\ & & & 1 & -1 \end{pmatrix} \theta, \quad (2.20)$$

since for $\theta = 0$ we have $A = G$ and $B = 0$. We now insert G^{-1} from (1.16) and observe that the matrices involved are, with the exception of two elements, equal to $\pm K$ of (2.17). We therefore approximate (2.20) by

$$\ddot{\theta} = -n^4 K^2 \theta. \quad (2.21)$$

This second order equation was integrated in IV.1 as a first order system

$$\begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix}' = \begin{pmatrix} 0 & I \\ -n^4 K^2 & 0 \end{pmatrix} \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix} = E \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix}. \quad (2.22)$$

By solving

$$\begin{pmatrix} 0 & I \\ -n^4 K^2 & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = \lambda \begin{pmatrix} y \\ z \end{pmatrix}, \quad (2.23)$$

we find that λ is an eigenvalue of E iff λ^2 is an eigenvalue of $-n^4 K^2$. Thus Formula (2.18) shows that the eigenvalues of E are situated on the imaginary axis between $-4n^2 i$ and $+4n^2 i$. We see from Fig. 2.2 that the stability domain of DOP853 covers the imaginary axis between approximately $-6i$ and $+6i$. Hence for stability we need $h \leq 1.5/n^2$ and the number of steps for the interval $0 \leq t \leq 5$ must be larger than $\approx 10n^2/3$. This, again, was observed in the numerical calculations (Table 1.2).

Automatic Stiffness Detection

Neither is perfect, but even an imperfect test can be quite useful,
as we can show from experience . . . (L.F. Shampine 1977)

Explicit codes applied to stiff problems are apparently not very efficient and the remaining part of the book will be devoted to the construction of more stable algorithms. In order to avoid that an explicit code waste too much effort when encountering stiffness (and to enable a switch to a more suitable method), it is important that the code be equipped with a cheap means of detecting stiffness. The analysis of the preceding subsection demonstrates that, whenever a nonstiff code encounters stiffness, the product of the step size with the dominant eigenvalue of the Jacobian lies near the border of the stability domain. We shall show two manners of exploiting this observation to detect stiffness.

Firstly, we adapt the ideas of Shampine & Hiebert (1977) to the Dormand & Prince method of order 5(4), given in Table II.5.2. The method possesses an error estimator $err_1 = y_1 - \hat{y}_1$ which, in the nonstiff situation, is $\mathcal{O}(h^5)$. However in the stiff case, when the method is working near the border of the stability domain S , the distance $d_1 = y_1 - y(x_0 + h)$ to the smooth solution is approximately $d_1 \approx R(hJ)d_0$, where J denotes the Jacobian of the system, $R(z)$ is the stability function of the method, and $d_0 = y_0 - y(x_0)$. Here we have neglected the local error for an initial value on the smooth solution $y(x)$. A similar formula, with R replaced by \hat{R} , holds for the embedded method. The error estimator satisfies $err_1 \approx E(hJ)d_0$ with $E(z) = R(z) - \hat{R}(z)$. The idea is now to search for a second error estimator \widetilde{err}_1 (with $\widetilde{err}_1 \approx \widetilde{E}(hJ)d_0$) such that

- i) $|\widetilde{E}(z)| \leq \theta |E(z)|$ on $\partial S \cap \mathbb{C}^-$ with a small $\theta < 1$;
- ii) $\widetilde{err}_1 = \mathcal{O}(h^2)$ for $h \rightarrow 0$.

Condition (i) implies that $\|\widetilde{err}_1\| < \|err_1\|$ when $h\lambda$ is near ∂S (the problem is possibly stiff), and condition (ii) will lead to $\|\widetilde{err}_1\| \gg \|err_1\|$ for step sizes which are determined by accuracy requirements (when the problem is not stiff). If $\|\widetilde{err}_1\| < \|err_1\|$ occurs several times in succession (say 15 times) then a stiff code might be more efficient.

For the construction of \widetilde{err}_1 we put $\widetilde{err}_1 = h(d_1 k_1 + d_2 k_2 + \dots + d_s k_s)$, where the $k_i = f(x_0 + c_i h, g_i)$ are the available function values of the method. The coefficients d_i are determined in such a way that

$$\sum_{i=1}^s d_i = 0, \quad \sum_{i=1}^s d_i c_i = 0.02 \quad (2.24)$$

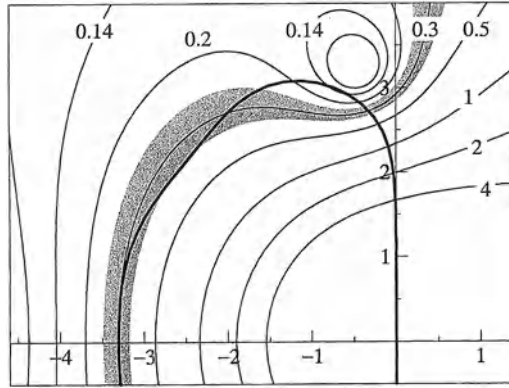
(so that (ii) holds) and that θ in (i) is minimized. A computer search gave values which have been rounded to

$$\begin{aligned} d_1 &= -0.08536, & d_2 &= 0.088, & d_3 &= -0.0096, \\ d_4 &= 0.0052, & d_5 &= 0.00576, & d_6 &= -0.004. \end{aligned} \quad (2.25)$$

The factor 0.02 in (2.24) has been chosen such that θ in (i) is close to 0.3 on

large parts of the border of S , but $|\tilde{E}(z)/E(z)|$ soon becomes larger than 1 if z approaches the origin.

In Fig. 2.4 we present the contour lines $|\tilde{E}(z)/E(z)| = \text{Const}$ ($\text{Const} = 4, 2, 1, 0.5, 0.3, 0.2, 0.14, 0.1$) together with the stability domain of the method. A numerical experiment is illustrated in Fig. 2.5. We applied the code DOPRI5 (see the Appendix to Volume I) to the van der Pol equation (1.5') with $\varepsilon = 0.003$. The upper picture shows the first component of the solution, the second picture displays the quotient $\|\tilde{err}_1\|/\|err_1\|$ for the three tolerances $Tol = 10^{-3}, 10^{-5}, 10^{-7}$. The last picture is a plot of $h|\lambda|/3.3$ where h is the current step size and λ the dominant eigenvalue of the Jacobian and 3.3 is the approximate distance of ∂S to the origin.



A second possibility for detecting stiffness is to estimate directly the dominant eigenvalue of the Jacobian of the problem. If v denotes an approximation to the corresponding eigenvector with $\|v\|$ sufficiently small then, by the mean value theorem,

$$|\lambda| \approx \frac{\|f(x, y + v) - f(x, y)\|}{\|v\|}$$

will be a good approximation to the leading eigenvalue. For the Dormand & Prince method (Table II.5.2) we have $c_6 = c_7 = 1$. Therefore, a natural choice is

$$\varrho = \frac{\|k_7 - k_6\|}{\|g_7 - g_6\|} \quad (2.26)$$

where $k_i = f(x_0 + c_i h, g_i)$ are the function values of the current step. Both values, $g_7 = y_1$ and g_6 , approximate the exact solution $y(x_0 + h)$ and it can be shown by Taylor expansion that $g_7 - g_6 = \mathcal{O}(h^3)$. This difference is thus sufficiently small, in general. The same argument also shows that $g_7 - g_6 = \tilde{E}(hJ)d_0$, where J is the Jacobian of the linearized differential equation and $\tilde{E}(z)$ is a polynomial with subdegree 4. Hence, $g_7 - g_6$ is essentially the vector obtained by 4 iterations of the power method applied to the matrix hJ . It will be a good approximation to the

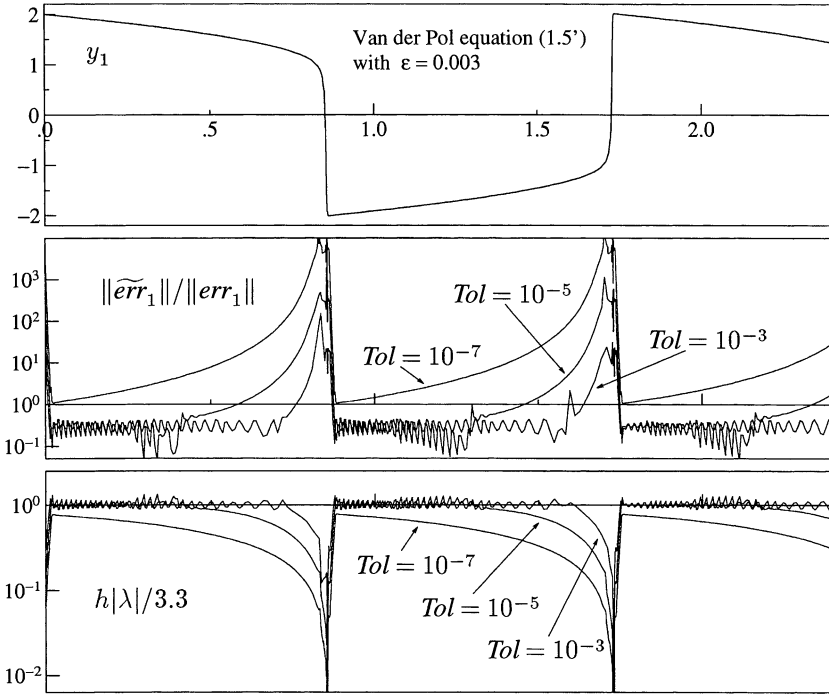


Fig. 2.5. Stiffness detection with DOPRI5

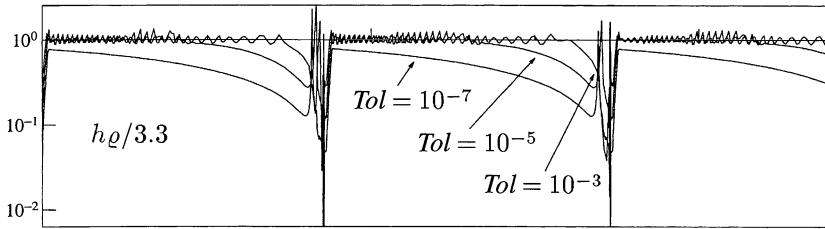


Fig. 2.6. Estimation of Lipschitz constant with DOPRI5

eigenvector corresponding to the leading eigenvalue. As in the above numerical experiment we applied the code DOPRI5 to the van der Pol equation (1.5') with $\varepsilon = 0.003$. Fig. 2.6 presents a plot of $h\rho/3.3$ where h is the current step size and ρ the estimate (2.26). This is in perfect agreement with the exact values $h|\lambda|/3.3$ (see third picture of Fig. 2.5).

Further numerical examples have shown that the estimate (2.26) also gives satisfactory approximations of $|\lambda|$ when the dominant eigenvalue λ is complex. However, if the argument of λ is needed too, one can extend the power method as proposed by Wilkinson (1965, page 579). This has been elaborated by Sottas (1984) and Robertson (1987).

The two techniques above allow us to detect the regions where the step size

is restricted by stability. In order to decide whether a stiff integrator will be more efficient, one has to compare the expense of both methods. Studies on this question have been undertaken in Petzold (1983), Sottas (1984) and Butcher (1990).

Step-Control Stability

We now come to the explanation of another phenomenon encountered in Sect. IV.1, that of the ragged behaviour of the step size (e.g. Fig. 1.4 or 1.8), a research initiated by G. Hall (1985/86) and continued by G. Hall & D.J. Higham (1988). Do there exist methods or stiff equations for which the step sizes h_n behave smoothly and no frequent step rejections appear?

We make a numerical study on the equation

$$\begin{aligned} y_1' &= -2000 (\cos x \cdot y_1 + \sin x \cdot y_2 + 1) & y_1(0) &= 1 \\ y_2' &= -2000 (-\sin x \cdot y_1 + \cos x \cdot y_2 + 1) & y_2(0) &= 0 \end{aligned} \quad (2.27)$$

for $0 \leq x \leq 1.57$, whose eigenvalues move slowly on a large circle from -2000 to $\pm 2000i$. If we apply Fehlberg's method RKF5(4) (Table II.5.1) in local extrapolation mode (i.e., we continue the integration with the higher order result) and DOPRI5 to this equation (with Euclidean error norm without scaling), we obtain the step size behaviour presented in Fig. 2.7. There all rejected steps are crossed out (3 rejected steps for RKF5(4) and 104 for DOPRI5).

In order to explain this behaviour, we consider for $y' = \lambda y$ (of course!) the numerical process

$$\begin{aligned} y_{n+1} &= R(h_n \lambda) y_n \\ err_n &= E(h_n \lambda) y_n \\ h_{n+1} &= h_n \cdot \left(\frac{Tol}{|err_n|} \right)^\alpha \end{aligned} \quad (2.28)$$

(where err_n is the estimated error, $E(z) = \hat{R}(z) - R(z)$, $\alpha = 1/(\hat{p} + 1)$ and \hat{p} is the order of \hat{R}) as a dynamical system whose fixed points and stability we have to study. A possible safety factor (“*fac*” of formula (4.13) of Sect. II.4) can easily be incorporated into Tol and does not affect the theory. The analysis simplifies if we introduce logarithms

$$\eta_n = \log |y_n|, \quad \chi_n = \log h_n \quad (2.29)$$

so that (2.28) becomes

$$\begin{aligned} \eta_{n+1} &= \log |R(e^{x_n} \lambda)| + \eta_n, \\ \chi_{n+1} &= \alpha \left(\gamma - \log |E(e^{x_n} \lambda)| - \eta_n \right) + \chi_n, \end{aligned} \quad (2.30)$$

where γ is a constant. This is now a map $\mathbb{R}^2 \rightarrow \mathbb{R}^2$. Its fixed point (η, χ) satisfies

$$|R(e^x \lambda)| = 1, \quad (2.31)$$

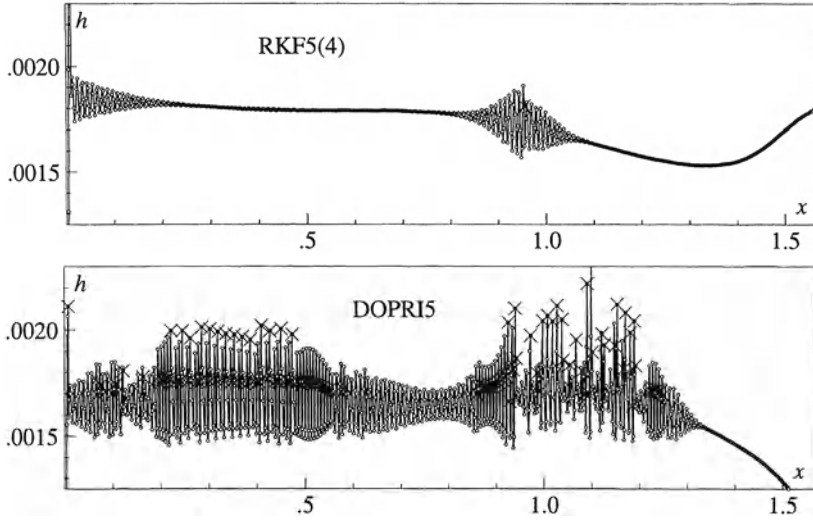


Fig. 2.7. Step sizes of RKF5(4) and DOPRI5 for (2.27)

which determines the step size e^x so that the point $z = e^x \lambda$ must be on the border of the stability domain. Further

$$\eta = \gamma - \log |E(z)|$$

determines η . Now the Jacobian of the map (2.30) at this fixed point becomes

$$C = \frac{\partial(\eta_{n+1}, \chi_{n+1})}{\partial(\eta_n, \chi_n)} = \begin{pmatrix} 1 & u \\ -\alpha & 1 - \alpha v \end{pmatrix} \quad \begin{aligned} u &= \operatorname{Re} \left(\frac{R'(z)}{R(z)} \cdot z \right) \\ v &= \operatorname{Re} \left(\frac{E'(z)}{E(z)} \cdot z \right). \end{aligned} \quad (2.32)$$

Proposition 2.3. *The step-control mechanism is stable for $h\lambda = z$ on the boundary of the stability domain if and only if the spectral radius of C in (2.32) satisfies*

$$\varrho(C) < 1.$$

We then call the method SC-stable at z . □

The matrix C is independent of the given differential equation and of the given tolerance. It is therefore a characteristic of the numerical method and the boundary of its stability domain. Let us study some methods of Sect. II.5.

a) RKF4(5) (Table 5.1), $\alpha = 1/5$:

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + \frac{z^5}{104}, \quad E(z) = \frac{z^5}{780} - \frac{z^6}{2080}. \quad (2.33)$$

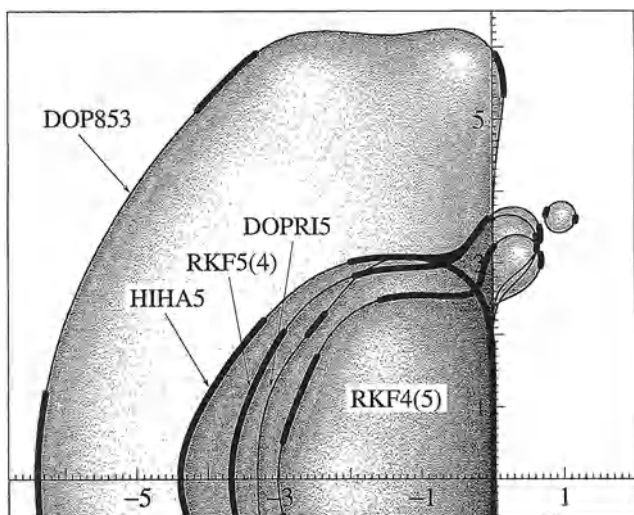


Fig. 2.8. Regions of step-control stability

b) DOPRI5 (Table 5.2), $\alpha = 1/5$:

$$R(z) = \text{see (2.13)}, \quad E(z) = \frac{97}{120000}z^5 - \frac{13}{40000}z^6 + \frac{1}{24000}z^7 \quad (2.34)$$

c) RKF5(4) (Table 5.1, with local extrapolation), $\alpha = 1/5$:

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + \frac{z^5}{120} + \frac{z^6}{2080}, \quad E(z) \text{ same as (2.33)}.$$

d) HIHA5 (Method of Higham & Hall, see Table 2.1 below), $\alpha = 1/5$:

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + \frac{z^5}{120} + \frac{z^6}{1440}, \quad (2.35)$$

$$E(z) = -\frac{1}{1200}z^5 + \frac{1}{2400}z^6 + \frac{1}{14400}z^7 \quad (2.36)$$

The corresponding stability domains are represented in Fig. 2.8. There, the regions of the boundary, for which $\rho(C) < 1$ is satisfied, are represented as **thick** lines. It can be observed that the phenomena of Fig. 2.7, as well as those of Sect. IV.1, are nicely verified.

DOP853. The step size control of the code DOP853 (Volume I) is slightly more complicated. It is based on a “stretched” error estimator (see Sect. II.10) and, for the test equation $y' = \lambda y$, it is equivalent to replacing $|E(z)|$ of (2.30) by

$$|E(z)| = \frac{|E_5(z)|^2}{\sqrt{|E_5(z)|^2 + 0.01 \cdot |E_3(z)|^2}}, \quad (2.37)$$

where $E_3(z) = \hat{R}_3(z) - R(z)$, $E_5(z) = \hat{R}_5(z) - R(z)$, and $\hat{R}_3(z)$, $\hat{R}_5(z)$ are the stability functions of third and fifth order embedded methods, respectively. The above analysis is still valid if the expression v of (2.32) is replaced by the derivative

Table 2.1. Method HIHA5 of Higham and Hall

0							
$\frac{2}{9}$	$\frac{2}{9}$						
$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{4}$					
$\frac{1}{2}$	$\frac{1}{8}$	0	$\frac{3}{8}$				
$\frac{3}{5}$	$\frac{91}{500}$	$-\frac{27}{100}$	$\frac{78}{125}$	$\frac{8}{125}$			
1	$-\frac{11}{20}$	$\frac{27}{20}$	$\frac{12}{5}$	$-\frac{36}{5}$	5		
1	$\frac{1}{12}$	0	$\frac{27}{32}$	$-\frac{4}{3}$	$\frac{125}{96}$	$\frac{5}{48}$	
y_1	$\frac{1}{12}$	0	$\frac{27}{32}$	$-\frac{4}{3}$	$\frac{125}{96}$	$\frac{5}{48}$	0
\widehat{y}_1	$\frac{2}{15}$	0	$\frac{27}{80}$	$-\frac{2}{15}$	$\frac{25}{48}$	$\frac{1}{24}$	$\frac{1}{10}$

of $\log |E(e^\chi \lambda)|$ with respect to χ , which is

$$v = 2v_5 - \frac{v_5 |E_5(z)|^2 + 0.01v_3 |E_3(z)|^2}{|E_5(z)|^2 + 0.01|E_3(z)|^2}, \quad (2.38)$$

where $v_5 = \operatorname{Re}(zE'_5(z)/E_5(z))$ and $v_3 = \operatorname{Re}(zE'_3(z)/E_3(z))$. Since $|E(z)| = \mathcal{O}(|z|^8)$ for $|z| \rightarrow 0$, we have to use the value $\alpha = 1/8$ in (2.32). The regions of SC -stability are shown in Fig. 2.8.

SC-Stable Dormand and Prince Pairs of Order 5. We see from Fig. 2.8 that the method DOPRI5 is not SC -stable at the intersection of the real axis with the boundary of the stability region. We are therefore interested in finding 5(4)-th order explicit Runge-Kutta pairs from the family of Dormand & Prince (1980) with larger regions of SC -stability.

Requiring the simplifying assumption (II.5.15), Algorithm 5.2 of Sect. II.5 yields a class of Runge-Kutta methods with c_3, c_4, c_5 as free parameters. Higham & Hall (1990) have made an extensive computer search for good choices of these parameters in order to have a reasonable size of the stability domain, large parts of SC -stability and a small 6th order error constant. It turned out that the larger one wants the region of SC -stability, the larger the error constant becomes. A compromise choice between Scylla and Charybdis, which in addition yields nice rational coefficients, is given by $c_3 = 1/3$, $c_4 = 1/2$ and $c_5 = 3/5$. This then leads to the method of Table 2.1 which has satisfactory stability properties as can be seen from Fig. 2.8.

A PI Step Size Control

We saw that it was an I-controller ... and a control-man knows that PI is always better than I ...

(K. Gustafsson, June 1990)

In 1986/87 two students of control theory attended a course of numerical analysis at the University of Lund. The outcome of this contact was the idea to resolve the above instability phenomena in stiff computations by using the concept of “PID control” (Gustafsson, Lundh & Söderlind 1988). The motivation for PID control, a classic in control theory (Callender, Hartree & Porter 1936) is as follows:

Suppose we have a continuous-time control problem where $\theta(t)$ is the departure, at time t , of a quantity to be controlled from its normal value. Then one might suppose that

$$\dot{\theta}(t) = C(t) - m\theta(t) \quad (2.39)$$

where $C(t)$ denotes the effect of the control and the term $-m\theta(t)$ represents a self-regulating effect such as “a vessel in a constant temperature bath”. The most simple assumption for the control would be

$$-\dot{C}(t) = n_1\theta(t) \quad (2.40)$$

which represents, say, a valve opened or closed in dependence of θ . The equations (2.39) and (2.40) together lead to

$$\ddot{\theta} + m\dot{\theta} + n_1\theta = 0 \quad (2.41)$$

which, for $n_1 > 0$, $m > 0$, is always stable. If, however, we assume (more realistically) that our system has some time-lag, we must replace (2.40) by

$$-\dot{C}(t) = n_1\theta(t - T) \quad (2.40')$$

and the stability of the process may be destroyed. This is precisely the same effect as the instability of Equation (17.6) of Sect. II.17 and is discussed similarly. In order to preserve stability, one might replace (2.40') by

$$-\dot{C}(t) = n_1\theta(t - T) + n_2\dot{\theta}(t - T) \quad (2.40'')$$

or even by

$$-\dot{C}(t) = n_1\theta(t - T) + n_2\dot{\theta}(t - T) + n_3\ddot{\theta}(t - T). \quad (2.40''')$$

Here, the first term on the right hand side represents the “Integral feedback” (I), the second term “Proportional feedback” (P) and the last term is the “Derivative feedback” (D). The P -term especially increases the constant m in (2.41), thus *adds extra friction* to the equation. It is thus natural to expect that the system becomes more stable. The precise tuning of the parameters n_1 , n_2 , n_3 is, however, a long task of analytic study and practical experience.

In order to adapt the continuous-time model (2.40'') to our situation, we replace

$$\begin{aligned} C(t) &\longleftrightarrow \log h_n \quad (\text{the “control variable”}) \\ \theta(t) &\longleftrightarrow \log |err_n| - \log Tol \quad (\text{the “deviation”}) \end{aligned}$$

and replace derivatives in t by differences. Then the formula (see (2.28))

$$h_{n+1} = h_n \cdot \left(\frac{Tol}{|err_n|} \right)^{n_1},$$

which is

$$-(\log h_{n+1} - \log h_n) = n_1 (\log |err_n| - \log Tol),$$

corresponds to (2.40'). The *PI*-control (2.40'') would read

$$\begin{aligned} -(\log h_{n+1} - \log h_n) &= n_1 (\log |err_n| - \log Tol) \\ &\quad + n_2 ((\log |err_n| - \log Tol) - (\log |err_{n-1}| - \log Tol)), \end{aligned}$$

or when resolved,

$$h_{n+1} = h_n \cdot \left(\frac{Tol}{|err_n|} \right)^{n_1} \left(\frac{|err_{n-1}|}{|err_n|} \right)^{n_2}. \quad (2.42)$$

In order to perform a *theoretical analysis* of this new algorithm we again choose the problem $y' = \lambda y$ and have as in (2.28)

$$y_{n+1} = R(h_n \lambda) y_n \quad (2.43a)$$

$$err_n = E(h_n \lambda) y_n \quad (2.43b)$$

$$\begin{aligned} h_{n+1} &= h_n \cdot \left(\frac{Tol}{|err_n|} \right)^{n_1} \left(\frac{|err_{n-1}|}{|err_n|} \right)^{n_2} \\ &= h_n \left(\frac{Tol}{|err_n|} \right)^\alpha \left(\frac{|err_{n-1}|}{Tol} \right)^\beta \end{aligned} \quad (2.43c)$$

where $\alpha = n_1 + n_2$, $\beta = n_2$. With the notation (2.29) this process becomes

$$\eta_{n+1} = \log |R(e^{\chi_n \lambda})| + \eta_n \quad (2.44)$$

$$\chi_{n+1} = \chi_n - \alpha \log |E(e^{\chi_n \lambda})| - \alpha \eta_n + \beta \log |E(e^{\chi_{n-1} \lambda})| + \beta \eta_{n-1} + \gamma$$

with some constant γ . This can be considered as a map $(\eta_n, \chi_n, \eta_{n-1}, \chi_{n-1}) \rightarrow (\eta_{n+1}, \chi_{n+1}, \eta_n, \chi_n)$. At a fixed point (η, χ) , which again satisfies (2.31), the Jacobian is given by

$$\tilde{C} = \frac{\partial(\eta_{n+1}, \chi_{n+1}, \eta_n, \chi_n)}{\partial(\eta_n, \chi_n, \eta_{n-1}, \chi_{n-1})} = \begin{pmatrix} 1 & u & 0 & 0 \\ -\alpha & 1 - \alpha v & \beta & \beta v \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad (2.45)$$

with u and v as in (2.32). A numerical study of the spectral radius $\varrho(\tilde{C})$ with $\alpha = 1/p$ (where p is the exponent of h of the leading term in the error estimator), $\beta = 0.08$ along the boundary of the stability domains of the above RK-methods shows an impressive improvement (see Fig. 2.9) as compared to the standard algorithm of Fig. 2.8. The only exception is DOP853, which becomes unstable close to the real

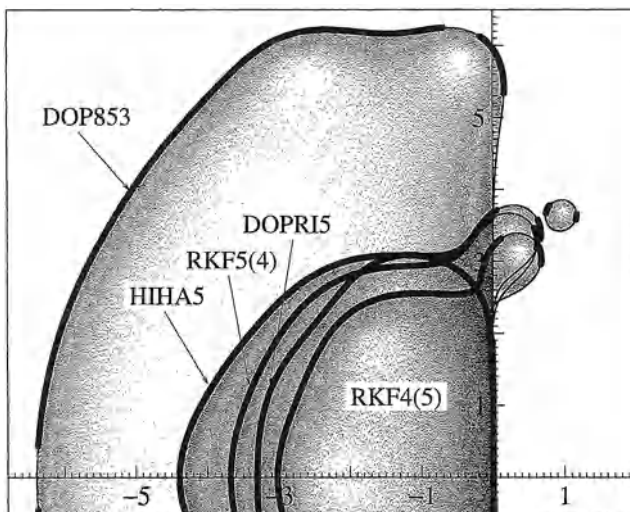


Fig. 2.9. Regions of step-control stability with stabilization factor $\beta = 0.08$

axis, whereas it was *SC*-stable for $\beta = 0$. For this method, the value $\beta = 0.04$ is more suitable.

The step size behaviour of DOPRI5 with the new strategy ($\beta = 0.13$) applied to the problem (1.6') is compared in Fig. 2.10 to the undamped step size control ($\beta = 0$). The improvement needs no comment. In order to make the difference clearly visible, we have chosen an extra-large tolerance $Atol = Rtol = 8 \cdot 10^{-2}$. With $\beta = 0.13$ the numerical solution becomes smooth in the time-direction. The zig-zag error in the x -direction represents the eigenvector corresponding to the largest eigenvalue of the Jacobian and its magnitude is below $Atol$.

Man sieht dass selbst der frömmste Mann
nicht allen Leuten gefallen kann.

(W. Busch, Kritik des Herzens 1874)

Study for small h . For the non-stiff case the new step size strategy may be slightly less efficient. In order to understand this, we assume that $|err_n| \approx Ch_n^p$ so that (2.43c) becomes

$$h_{n+1} = h_n \left(\frac{Tol}{Ch_n^p} \right)^\alpha \left(\frac{Ch_{n-1}^p}{Tol} \right)^\beta \quad (2.46)$$

or, by taking logarithms,

$$\log h_{n+1} + (p\alpha - 1) \log h_n - p\beta \log h_{n-1} = (\alpha - \beta) \log \left(\frac{Tol}{C} \right).$$

This is a linear difference equation with characteristic equation

$$\lambda^2 + (p\alpha - 1)\lambda - p\beta = 0, \quad (2.47)$$

the roots of which govern the response of the system to variations in C . Obviously, the choice $\alpha = 1/p$ and $\beta = 0$ would be most perfect by making both roots equal to

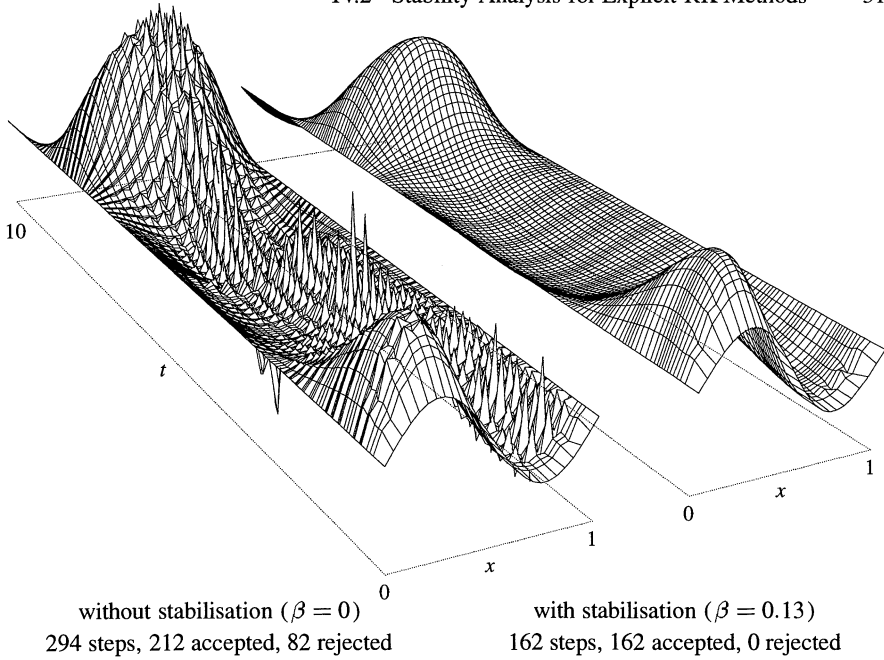


Fig. 2.10. Numerical solution of (1.6') with $Tol = 8 \cdot 10^{-2}$

zero; but this is just the classical step size control. We therefore have to compromise by choosing α and β such that (2.45) remains stable for large parts of the stability boundary and at the same time keeping the roots of (2.47) significantly smaller than one. A fairly good choice, found by Gustafsson (1991) after some numerical computations, is

$$\alpha \approx 0.7/p, \quad \beta \approx 0.4/p. \quad (2.48)$$

Stabilized Explicit Runge-Kutta Methods

For many problems, usually not very stiff, of large dimension, and with eigenvalues known to lie in a certain region, explicit methods with large stability domains can be very efficient. We consider here methods with extended stability domains along the negative real axis, which are, therefore, especially suited for the time integration of systems of parabolic PDEs. An excellent survey article with additional details and references is Verwer (1996).

Our problem is to find, for a given s , a polynomial of the form $R(z) = 1 + z + a_2 z^2 + \dots + a_s z^s$ such that the corresponding stability domain is, in the direction of the negative axis, as large as possible. The main ingredient for these methods are the Chebyshev polynomials (Chebyshev 1854)

$$T_s(x) = \cos(s \arccos x) \quad (2.49)$$

or

$$T_s(x) = 2xT_{s-1}(x) - T_{s-2}(x), \quad T_0(x) = 1, \quad T_1(x) = x \quad (2.49')$$

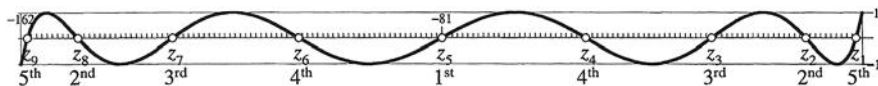


Fig. 2.11. Shifted Chebyshev polynomial $T_9(1 + z/81)$ and its zeros

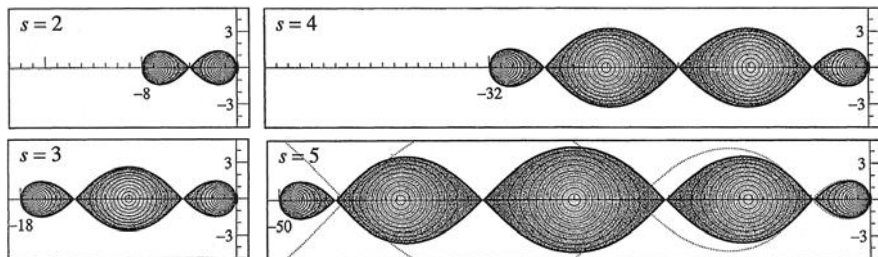


Fig. 2.12. Stability domains for shifted Chebyshev polynomials ($s = 2, 3, 4, 5$) (dots represent limiting case $s \rightarrow \infty$, see Exercise 8 below)

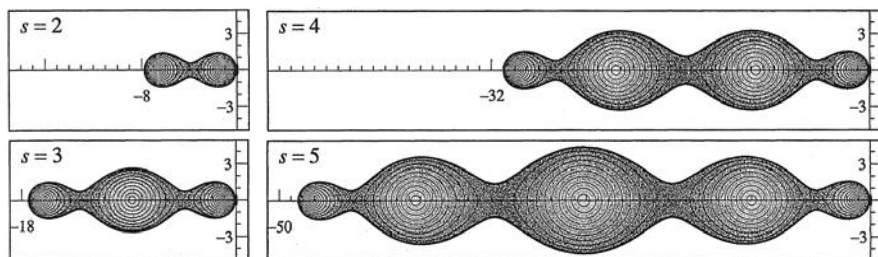


Fig. 2.13. Stability domains for *damped* Chebyshev stability functions, $\varepsilon = 0.05$

which remain for $-1 \leq x \leq 1$ between -1 and $+1$ and among these polynomials have the largest possible derivative $T'_s(1) = s^2$ (A.A. Markov 1890). Therefore, one must set (Saul'ev 1960, Saul'ev's postgraduate student Yuan Chzao Din 1958, Franklin 1959, Guillou & Lago 1961)

$$R_s(z) = T_s(1 + z/s^2) \quad (2.50)$$

so that $R_s(0) = 1$, $R'_s(0) = 1$, and $|R_s(z)| \leq 1$ for $-2s^2 \leq z \leq 0$ (see Fig. 2.11). In particular we have

$$\begin{aligned} R_1(z) &= 1 + z \\ R_2(z) &= 1 + z + \frac{1}{8}z^2 \\ R_3(z) &= 1 + z + \frac{4}{27}z^2 + \frac{4}{729}z^3 \\ R_4(z) &= 1 + z + \frac{5}{32}z^2 + \frac{1}{128}z^3 + \frac{1}{8192}z^4 \\ R_5(z) &= 1 + z + \frac{4}{25}z^2 + \frac{28}{3125}z^3 + \frac{16}{78125}z^4 + \frac{16}{9765625}z^5. \end{aligned} \quad (2.50')$$

whose stability domains are represented in Fig. 2.12.

Damping. In the points where $T_s(1 + z/s^2) = \pm 1$, there is no damping at all of the higher frequencies and the stability domain has zero width. We therefore choose a small $\varepsilon > 0$, say $\varepsilon = 0.05$, and put (already suggested by Guillou & Lago 1961)

$$R_s(z) = \frac{1}{T_s(w_0)} T_s(w_0 + w_1 z), \quad w_0 = 1 + \frac{\varepsilon}{s^2}, \quad w_1 = \frac{T_s(w_0)}{T'_s(w_0)}. \quad (2.51)$$

These polynomials oscillate between approximately $1 - \varepsilon$ and $-1 + \varepsilon$ and again satisfy $R_s(z) = 1 + z + \mathcal{O}(z^2)$. The stability domains become a bit shorter (by $(4\varepsilon/3)s^2$), but the boundary is in a safe distance from the real axis (see Fig. 2.13).

Lebedev's Realization. Our next problem is to find Runge-Kutta methods which realize these stability polynomials. A first idea, mentioned by Saul'ev (1960) and Guillou & Lago (1961), is to write

$$R_s(z) = \prod_{i=1}^s (1 + \delta_i z) \quad \text{where} \quad \delta_i = -\frac{1}{z_i}, \quad z_i \text{ roots of } R(z) \quad (2.52)$$

and to represent the RK method as the *composition* of explicit Euler steps

$$g_0 := y_0, \quad g_i := g_{i-1} + h\delta_i f(g_{i-1}), \quad (i = 1, 2, \dots, s), \quad y_1 := g_s. \quad (2.53)$$

A disadvantage here is the fact that for the first of these roots, which in absolute value is much smaller than the others, we shall have a very large Euler step, which is surely not good. Lebedev's idea (Lebedev 1989, 1994) is therefore to group the roots symmetrically two-by-two together and to represent the corresponding quadratic factor

$$(1 + \delta_i z)(1 + \delta'_i z) = (1 + 2\alpha_i z + \beta_i z^2) \quad (2.54)$$

by a two-stage scheme

$$\begin{aligned} g_i &:= g_{i-1} + h\alpha_i f(g_{i-1}) \\ g_{i+1}^* &:= g_i + h\alpha_i f(g_i) \\ g_{i+1} &:= g_{i+1}^* - h\alpha_i \gamma_i (f(g_i) - f(g_{i-1})) \\ &= g_{i+1}^* - \gamma_i ((g_{i+1}^* - g_i) - (g_i - g_{i-1})) \end{aligned} \quad (2.55)$$

which produces (2.54) if $\beta_i = \alpha_i^2(1 - \gamma_i)$. This halves nearly the largest Euler step size and allows also complex conjugate pairs of roots. The expression $(g_{i+1}^* - g_i) - (g_i - g_{i-1}) \approx h^2 \alpha_i^2 y''$ can be used for error estimations and step size selections. For odd s , there remains one single root which gives rise to an Euler step (2.53).

Best Ordering. Some attention is now necessary for the decision in which order the roots shall be used (Lebedev & Finogenov 1976). This is done by two requirements: firstly, the quantities

$$S_j = \max_z |1 + \delta_1 z| \prod_{i=1}^j |1 + 2\alpha_i z + \beta_i z^2|,$$

which express the stability of the internal stages, must be ≤ 1 (here, the max is taken over real z in the stability interval of the method). Secondly, the quantities

$$Q_j = \max_z \prod_{i=j+1}^s |1 + 2\alpha_i z + \beta_i z^2|,$$

which describe the propagation of rounding errors, must be as small as possible. These conditions, evaluated numerically for the case $s = 9$, lead to the ordering indicated in Fig. 2.11.

Second Order Methods. If the stability polynomial is a second order approximation to e^z , i.e., if

$$R_s(z) = 1 + z + \frac{z^2}{2} + a_3 z^3 + \dots + a_s z^s \quad (2.56)$$

then it can be seen from (2.8) that any corresponding Runge-Kutta scheme is also of second order for nonlinear problems. Analytic expressions, in terms of an elliptic integral, for such optimal polynomials have been obtained by Lebedev & Medovikov (1994). Their stability region reaches to $-0.821842 \cdot s^2$ for $s \gg 1$. Their practical computation is usually done numerically (Remez 1957, Lebedev 1995). For example, in the case $s = 9$ and for a damping factor $\varepsilon = 0.015$, we obtain the roots

$$\begin{aligned} z_9 &= -64.64238389, & z_8 &= -60.67479347, & z_7 &= -53.21695488, \\ z_6 &= -43.16527010, & z_5 &= -31.72471699, & z_4 &= -20.25474163, \\ z_3 &= -10.05545938, & z_{2,1} &= -1.30596166 \pm i \cdot 1.34047517 \end{aligned} \quad (2.57)$$

The corresponding stability polynomials, which are stable for $-65.15 \leq z \leq 0$, the stability domain, and the best ordering are shown in Fig. 2.14. We see that we now have a pair of complex roots.

Lebedev's computer code, called DUMKA, incorporates the formulas of the above type with automatic selection of h and s in a wide range.

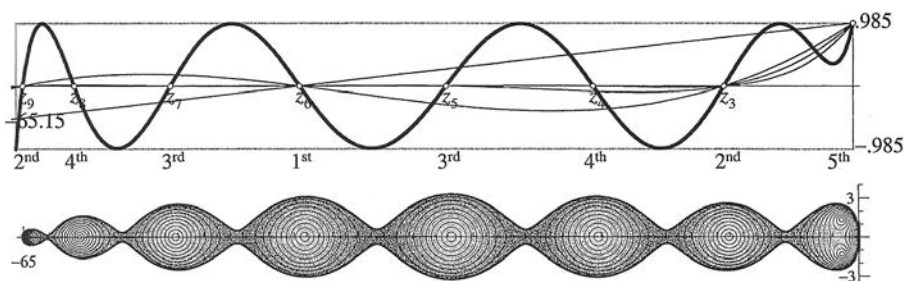


Fig. 2.14. Second order Zolotarev approximation with stability domain

Numerical Example. As an illustration, the method corresponding to (2.55) and (2.57) has been applied to problem (1.6'). Theory predicts stability for approximately $h \leq 65.15/135 = 0.4826$. The leftmost picture of Fig. 2.15 is computed

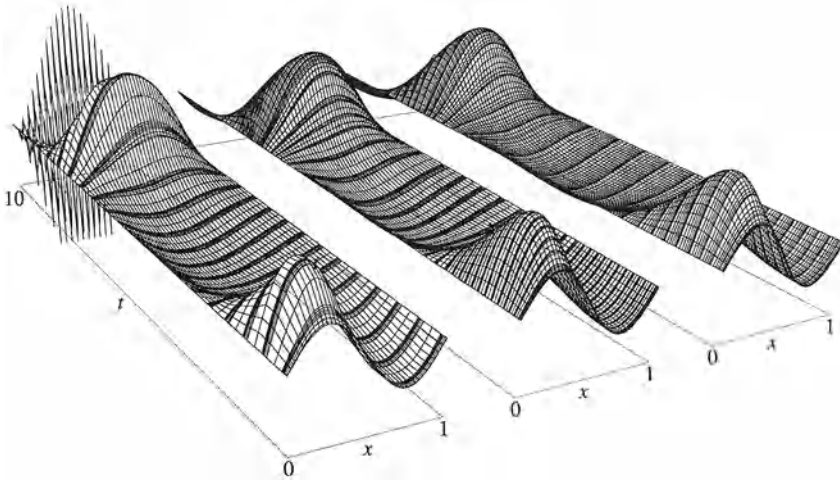


Fig. 2.15. Problem (1.6'): Lebedev9, $h = 0.48865$ (left), DUMKA (middle), RKC (right) (all internal stages drawn)

with $h = 0.48865$, which is a little too large and produces instability. The middle picture is produced by the code DUMKA with $Tol = 3 \cdot 10^{-3}$.

The Approach of van der Houwen & Sommeijer. An elegant idea for a second realization has been found by van der Houwen & Sommeijer (1980): apply scaled and shifted Chebyshev polynomials and use the three-term recursion formula (2.49') for defining the internal stages. We therefore, following Bakker (1973), set

$$R_s(z) = a_s + b_s T_s(w_0 + w_1 z) \quad w_0 = 1 + \varepsilon/s^2, \quad \varepsilon \approx 0.15. \quad (2.58)$$

The conditions for second order

$$R_s(0) = 1, \quad R'_s(0) = 1, \quad R''_s(0) = 1$$

lead to

$$w_1 = \frac{T'_s(w_0)}{T''_s(w_0)}, \quad b_s = \frac{T''_s(w_0)}{(T'_s(w_0))^2}, \quad a_s = 1 - b_s T_s(w_0), \quad (2.59)$$

with damping $a_s + b_s \approx 1 - \varepsilon/3$ (see Ex. 9). We now put for the internal stages

$$R_j(z) = a_j + b_j T_j(w_0 + w_1 z) \quad j = 0, 1, \dots, s-1. \quad (2.60)$$

It has been discovered by Sommeijer (see Sommeijer & Verwer 1980), that these $R_j(z)$ can, for $j \geq 2$, be approximations of second order at certain points $x_0 + c_j h$ if

$$R_j(0) = 1, \quad R'_j(0) = c_j, \quad R''_j(0) = c_j^2 \quad (2.61)$$

which gives

$$R_j(z) - 1 = b_j (T_j(w_0 + w_1 z) - T_j(w_0)), \quad b_j = \frac{T''_j(w_0)}{(T'_j(w_0))^2}. \quad (2.62)$$

The three-term recurrence relation (2.49') now leads to

$$R_j(z) - 1 = \mu_j(R_{j-1}(z) - 1) + \nu_j(R_{j-2}(z) - 1) + \kappa_j \cdot z \cdot (R_{j-1}(z) - a_{j-1})$$

where

$$\mu_j = \frac{2b_j w_0}{b_{j-1}}, \quad \nu_j = \frac{-b_j}{b_{j-2}}, \quad \kappa_j = \frac{2b_j w_1}{b_{j-1}}, \quad j = 2, 3, \dots, s. \quad (2.63)$$

This formula allows, in the case of a nonlinear differential system, to define the scheme

$$\begin{aligned} g_0 - y_0 &= 0, \\ g_1 - y_0 &= \kappa_1 h f(g_0), \\ g_j - y_0 &= \mu_j(g_{j-1} - y_0) + \nu_j(g_{j-2} - y_0) + \kappa_j h f(g_{j-1}) - a_{j-1} \kappa_j h f(g_0), \end{aligned} \quad (2.64)$$

which, being of second order for $y' = \lambda y$, is of second order for nonlinear equations too (again because of (2.8)). For $j = 1$ only first order is possible and κ_1 can be chosen freely. Sommeijer & Verwer (1980) suggest to put

$$b_0 = b_2, \quad b_1 = b_2 \quad \text{which gives} \quad \kappa_1 = c_1 = \frac{c_2}{T'_2(w_0)} \approx \frac{c_2}{4}.$$

Fig. 2.16 shows, for $s = 9$ as usual, the functions $R_s(z)$ and $R_j(z)$, $j = 2, \dots, s - 1$ together with the stability domain of $R_s(z)$ (the “Venus of Willendorf”) in exactly the same frame as Lebedev’s Zolotarev polynomial of Fig. 2.14. We see that the stability domain becomes a little shorter, but we have closed analytic expressions and a smoother behaviour of the c_i ’s (see Fig. 2.15, right). All internal stages satisfy $|R_j(z)| \leq 1$, and the method can be seen to possess a satisfactory numerical stability (see Verwer, Hundsdorfer & Sommeijer 1990). The above formulas have been implemented in a research code RKC (“Runge-Kutta-Chebyshev”) by Sommeijer (1991). As can be seen from Fig. 2.15, it performs well for equation (1.6’). More numerical results shall be reported in Sect. IV.10.

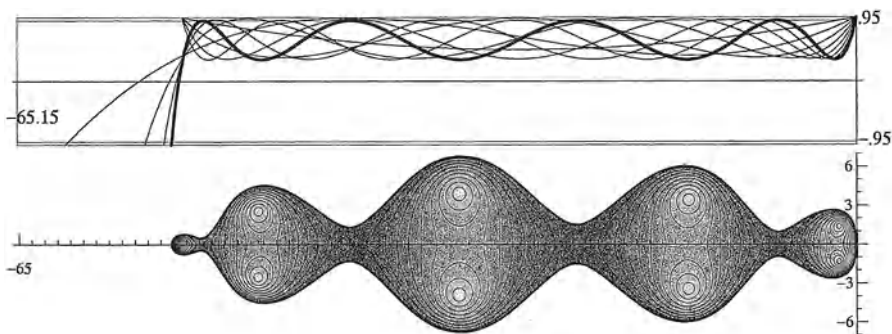


Fig. 2.16. Stability function and domain for RKC method, $s = 9$, $\varepsilon = 0.15$

Combined Approach of Abdulle & Medovikov. Recent research and a code ROCK4 are presented in: A. Abdulle, *Fourth order Chebyshev methods with recurrence relation*, to appear in SIAM J. Sci. Comput. 2002.

Exercises

1. Prove that Runge-Kutta methods are invariant under linear transformations $y = Tz$ (i.e., if one applies the method to $y' = f(x, y)$ and to $z' = T^{-1}f(x, Tz)$ with initial values satisfying $y_0 = Tz_0$, then we have $y_1 = Tz_1$).
2. Consider the differential equation $y' = Ay$ and a numerical solution given by $y_{n+1} = R(hA)y_n$. Suppose that $R(z)$ is A -stable, i.e., it satisfies

$$|R(z)| \leq 1 \quad \text{for} \quad \operatorname{Re} z \leq 0,$$

and show, by transforming A to Jordan canonical form, that

- a) if $y' = Ay$ is stable, then $\{y_n\}$ is bounded;
 - b) if $y' = Ay$ is asymptotically stable, then $y_n \rightarrow 0$ for $n \rightarrow \infty$.
3. (Optimal stability for hyperbolic problems, van der Houwen (1968), (1977), p.99): Given m , find a polynomial $R_m(z) = 1 + z + \dots$ of degree $m + 1$ such that $|R(iy)| \leq 1$ for $-\beta \leq y \leq \beta$ with β as large as possible.

Result. The solution (Sonneveld & van Leer 1985) is given by

$$R_m(z) = \frac{1}{2}V_{m-1}(\zeta) + V_m(\zeta) + \frac{1}{2}V_{m+1}(\zeta), \quad \zeta = \frac{z}{m} \quad (2.65)$$

where $V_m(\zeta) = i^m T_m(\zeta/i)$ are the Chebyshev polynomials with positive coefficients. $R_m(iy)$ is stable for $-m \leq y \leq m$. The first R_m are (see Abramowitz & Stegun, p. 795)

$$\begin{aligned} R_1(z) &= 1 + \zeta + \zeta^2 & \zeta &= \frac{z}{m} \\ R_2(z) &= 1 + 2\zeta + 2\zeta^2 + 2\zeta^3 \\ R_3(z) &= 1 + 3\zeta + 5\zeta^2 + 4\zeta^3 + 4\zeta^4 \\ R_4(z) &= 1 + 4\zeta + 8\zeta^2 + 12\zeta^3 + 8\zeta^4 + 8\zeta^5 \end{aligned} \quad (2.66)$$

Similar as for Chebyshev polynomials, they satisfy the recurrence relation $R_{m+1} = 2\zeta R_m + R_{m-1}$ ($m \geq 2$). Their stability domains are given in Fig. 2.17.

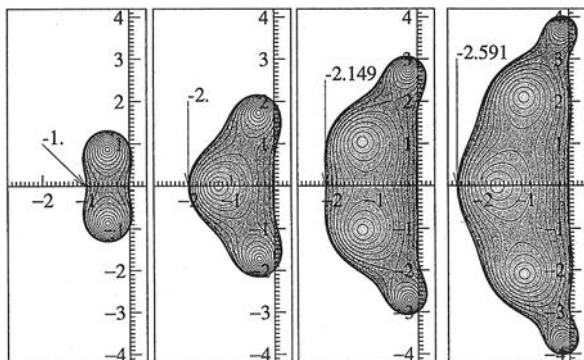


Fig. 2.17. Stability domains for hyperbolic approximations

4. Linearize the rope equation (1.24) in the neighbourhood of $\theta = \dot{\theta} = 0$ and make a stability analysis. Re-obtain Lagrange's equation (I.6.2) from the linearized equation with the coordinate transformation

$$y = \begin{pmatrix} 1 \\ 1 & 1 \\ 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \theta, \quad \theta = \begin{pmatrix} 1 \\ -1 & 1 & 1 \\ & -1 & 1 & \ddots \\ & & \ddots & \ddots \end{pmatrix} y.$$

5. Fig. 2.18 shows the numerical results of the classical 4th order Runge-Kutta method with equidistant steps over $0 \leq t \leq 5$ for the beam problem (1.7)-(1.20) with $n = 8$. Explain the result with the help of Fig. 2.1.

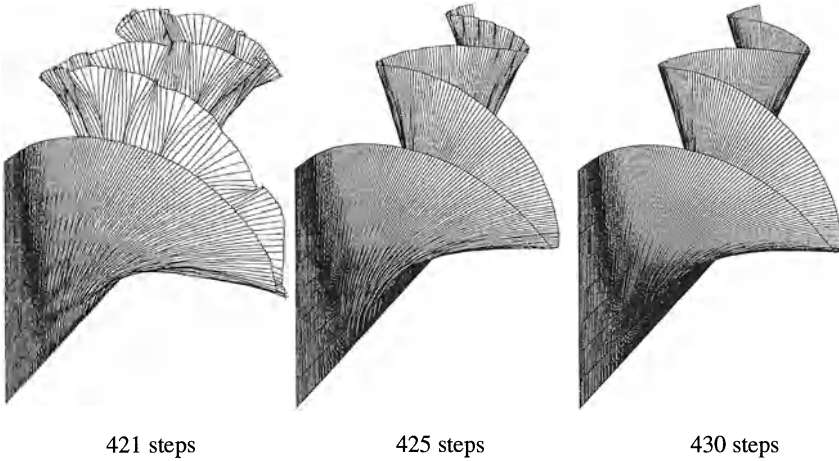


Fig. 2.18. Classical Runge-Kutta method (constant step sizes) on the beam problem

6. For the example of Exercise 5, the explicit Euler method, although converging for $h \rightarrow 0$, is *never* stable (see Fig. 2.19). Why?
7. Let λ be an eigenvalue of the two-dimensional left upper submatrix of \tilde{C} in (2.45) (matrix C of (2.32)) and denote its analytic continuation as eigenvalue of \tilde{C} by $\lambda(\beta)$. Prove that
- a) If $\operatorname{Re} \lambda \neq 0$, then for some $y \in \mathbb{R}$

$$\lambda(\beta) = \lambda \cdot \left(1 - \frac{\beta}{\alpha} (1 - \operatorname{Re} \lambda) + i\beta y + \mathcal{O}(\beta^2) \right).$$

This shows that $|\lambda(\beta)| < |\lambda|$ for small $\beta > 0$ if $\operatorname{Re} \lambda < 1$.

- b) If λ and μ are two distinct real eigenvalues of the above mentioned submatrix, then

$$\lambda(\beta) = \lambda \cdot \left(1 - \frac{\beta}{\alpha} \left(1 - \frac{1}{\lambda} \right)^2 \frac{1}{\lambda - \mu} + \mathcal{O}(\beta^2) \right).$$

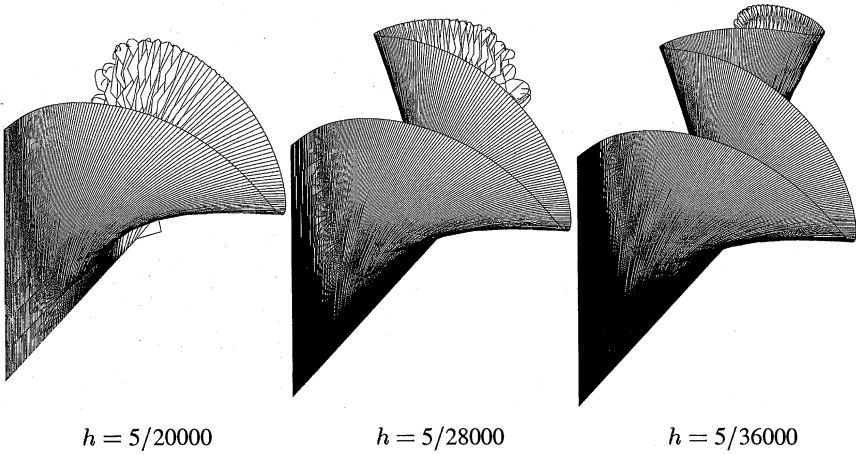


Fig. 2.19. Explicit Euler on the beam problem (every 50th step drawn)

Hint. Write the characteristic polynomial of \tilde{C} in the form

$$\det(\lambda I - \tilde{C}) = \lambda(\lambda p(\lambda) + \beta q(\lambda)),$$

where $p(\lambda) = \det(\lambda I - C)$ is the characteristic polynomial of C , and differentiate with respect to β .

8. Show that for the Chebyshev stability functions (2.50) we have

$$\lim_{s \rightarrow \infty} R_s(z) = \cos(\sqrt{-2z}).$$

Hint. Insert $\arccos(1 - x^2/2) \approx x$ into (2.49) and (2.50). The corresponding stability domain is indicated by dotted lines in the last picture of Fig. 2.12.

9. Show (for example with the help of (2.49')) that for the Chebyshev polynomials

$$T'_s(1) = s^2, \quad T''_s(1) = \frac{s^2(s^2 - 1)}{3}$$

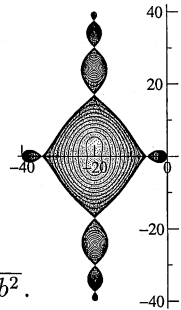
and obtain asymptotic values (for $\varepsilon \rightarrow 0$) for w_1 , b_s , a_s , the damping factor and the stability interval of the Bakker polynomials (2.58).

10. (Cross-shaped stability domains). For $-1 \leq \varphi \leq 1$ we put $z = -b \pm \sqrt{a(\varphi - 1) + b^2}$, so that z moves on a cross $-2b \leq z \leq 0$ and $z = -b \pm iy$. Thus (an idea of Lebedev)

$$R_{2s}(z) = T_s(\varphi(z))$$

is a stability function for eigenvalues on crosses (as, e.g., for the PLATE problem). Determine a in dependence of b from the condition $R'(0) = 1$ and find the maximal value for y .

Result. $R_{2s}(z) = T_s(1 + z/s^2 + z^2/(2bs^2))$; $y_{\max} = \sqrt{4bs^2 - b^2}$.



IV.3 Stability Function of Implicit RK-Methods

I didn't like all these "strong", "perfect", "absolute", "generalized", "super", "hyper", "complete" and so on in mathematical definitions, I wanted something neutral; and having been impressed by David Young's "property A", I chose the term "A-stable".

(G. Dahlquist, in 1979)

There are at least two ways to combat stiffness. One is to design a better computer, the other, to design a better algorithm.

(H. Lomax in Aiken 1985)

Methods are called *A*-stable if there are no stability restrictions for $y' = \lambda y$, $\operatorname{Re} \lambda < 0$ and $h > 0$. This concept was introduced by Dahlquist (1963) for linear multi-step methods, but also applied to Runge-Kutta processes. Ehle (1968) and Axelson (1969) then independently investigated the *A*-stability of implicit Runge-Kutta methods and proposed new classes of *A*-stable methods. A nice paper of Wright (1970) studied collocation methods.

The Stability Function

We start with the implicit Euler method. This method, $y_1 = y_0 + hf(x_1, y_1)$, applied to Dahlquist's equation $y' = \lambda y$ becomes $y_1 = y_0 + h\lambda y_1$ which, after solving for y_1 , gives

$$y_1 = R(h\lambda) y_0 \quad \text{with} \quad R(z) = \frac{1}{1-z}.$$

This time, the stability domain is the *exterior* of the circle with radius 1 and centre $+1$. The stability domain thus covers the *entire* negative half-plane and a large part of the positive half-plane as well. The implicit Euler method is *very* stable.

Proposition 3.1. *The s -stage implicit Runge-Kutta method*

$$g_i = y_0 + h \sum_{j=1}^s a_{ij} f(x_0 + c_j h, g_j) \quad i = 1, \dots, s \quad (3.1a)$$

$$y_1 = y_0 + h \sum_{j=1}^s b_j f(x_0 + c_j h, g_j) \quad (3.1b)$$

applied to $y' = \lambda y$ yields $y_1 = R(h\lambda)y_0$ with

$$R(z) = 1 + zb^T(I - zA)^{-1}\mathbb{1}, \quad (3.2)$$

where $b^T = (b_1, \dots, b_s)$, $A = (a_{ij})_{i,j=1}^s$, $\mathbb{1} = (1, \dots, 1)^T$.

Remark. As in Definition 2.1, $R(z)$ is called the *stability function* of Method (3.1).

Proof. Equation (3.1a) with $f(x, y) = \lambda y$, $z = h\lambda$ becomes a linear system for the computation of g_1, \dots, g_s . Solving this and inserting into (3.1b) leads to (3.2). \square

Another useful formula for $R(z)$ is the following (Stetter 1973, Scherer 1979):

Proposition 3.2. *The stability function of (3.1) satisfies*

$$R(z) = \frac{\det(I - zA + z\mathbb{1}b^T)}{\det(I - zA)}. \quad (3.3)$$

Proof. Applying (3.1) to (2.9) yields the linear system

$$\begin{pmatrix} I - zA & 0 \\ -zb^T & 1 \end{pmatrix} \begin{pmatrix} g \\ y_1 \end{pmatrix} = y_0 \begin{pmatrix} \mathbb{1} \\ 1 \end{pmatrix}.$$

Cramer's rule (Cramer 1750) implies that the denominator of $R(z)$ is $\det(I - zA)$, and its numerator

$$\det \begin{pmatrix} I - zA & \mathbb{1} \\ -zb^T & 1 \end{pmatrix} = \det \begin{pmatrix} I - zA + z\mathbb{1}b^T & 0 \\ -zb^T & 1 \end{pmatrix} = \det(I - zA + z\mathbb{1}b^T). \quad \square$$

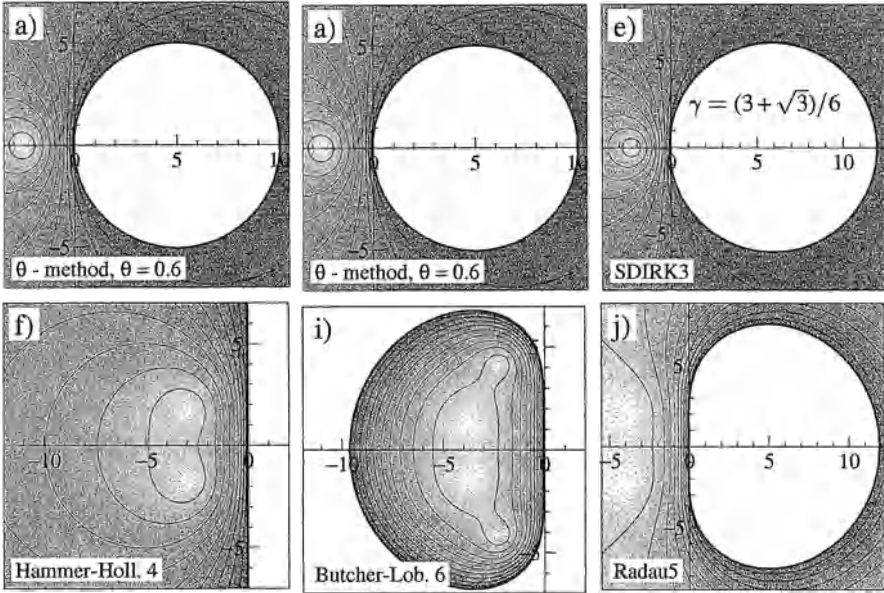


Fig. 3.1. Stability domains for implicit Runge-Kutta methods

The stability functions for the methods of Sect. II.7 are presented in Table 3.1. The corresponding stability domains are displayed in Fig. 3.1.

We see that for implicit methods $R(z)$ becomes a rational function with numerator and denominator of degree $\leq s$. We write

$$R(z) = \frac{P(z)}{Q(z)}, \quad \deg P = k, \quad \deg Q = j. \quad (3.4)$$

Table 3.1. Stability functions for implicit Runge-Kutta methods of Sect. II.7

	Method	$R(z)$
a)	θ -method (II.7.2)	$\frac{1+z(1-\theta)}{1-z\theta}$
b)	implicit Euler (II.7.3)	$\frac{1}{1-z}$
c)	implicit midpoint (II.7.4) } trapezoidal rule (II.7.5) }	$\frac{1+z/2}{1-z/2}$
d)	Hammer-Hollingsworth (II.7.6)	$\frac{1+4z/6+z^2/6}{1-z/3}$
e)	SDIRK order 3 (Table II.7.2)	$\frac{1+z(1-2\gamma)+z^2(1/2-2\gamma+\gamma^2)}{(1-\gamma z)^2}$
f)	Hammer-Hollingsw. 4 (Table II.7.3) } Lobatto IIIA, order 4 (Table II.7.7) }	$\frac{1+z/2+z^2/12}{1-z/2+z^2/12}$
g)	Kuntzm.-Butcher 6 (Table II.7.4)	$\frac{1+z/2+z^2/10+z^3/120}{1-z/2+z^2/10-z^3/120}$
h)	Butcher's Lobatto 4 (Table II.7.6)	$\frac{1+3z/4+z^2/4+z^3/24}{1-z/4}$
i)	Butcher's Lobatto 6 (Table II.7.6)	$\frac{1+2z/3+z^2/5+z^3/30+z^4/360}{1-z/3+z^2/30}$
j)	Radau IIA, order 5 (Table II.7.7)	$\frac{1+2z/5+z^2/20}{1-3z/5+3z^2/20-z^3/60}$

If the method is of order p , then

$$e^z - R(z) = Cz^{p+1} + \mathcal{O}(z^{p+2}) \quad \text{for } z \rightarrow 0 \quad (3.5)$$

(see Theorem 2.2). The constant C is usually $\neq 0$. If not, we increase p in (3.5) until C becomes $\neq 0$. We then call $R(z)$ a *rational approximation to e^z of order p* and C its *error constant*.

A-Stability

We observe that some methods are stable on the entire left half-plane \mathbb{C}^- . This is precisely the set of eigenvalues, where the *exact* solution of (2.9) is stable too (Sect. I.13, Theorem 13.1). A desirable property for a numerical method is that it preserves this stability property.

Definition 3.3 (Dahlquist 1963). A method, whose stability domain satisfies

$$S \supset \mathbb{C}^- = \{z; \operatorname{Re} z \leq 0\},$$

is called *A-stable*.

A Runge-Kutta method with (3.4) as stability function is A -stable if and only if

$$|R(iy)| \leq 1 \quad \text{for all real } y \quad (3.6)$$

and

$$R(z) \quad \text{is analytic for } \operatorname{Re} z < 0. \quad (3.7)$$

This follows from the maximum principle applied to \mathbb{C}^- . By a slight abuse of language, we also call $R(z)$ A -stable in this case (or, as many authors say, “ A -acceptable”, Ehle 1968).

Condition (3.6) alone means stability on the imaginary axis and may be called I -stability. It is equivalent to the fact that the polynomial

$$E(y) = |Q(iy)|^2 - |P(iy)|^2 = Q(iy)Q(-iy) - P(iy)P(-iy) \quad (3.8)$$

satisfies

$$E(y) \geq 0 \quad \text{for all } y \in \mathbb{R}. \quad (3.9)$$

Proposition 3.4. $E(y)$, defined by (3.8), is an even polynomial of degree $\leq 2 \max(\deg P, \deg Q)$. If $R(z)$ is an approximation of order p , then

$$E(y) = \mathcal{O}(y^{p+1}) \quad \text{for } y \rightarrow 0.$$

Proof. Taking absolute values in (3.5) gives

$$|e^z| - \frac{|P(z)|}{|Q(z)|} = \mathcal{O}(z^{p+1}).$$

Putting $z = iy$ and using $|e^{iy}| = 1$ leads to

$$|Q(iy)| - |P(iy)| = \mathcal{O}(y^{p+1}).$$

The result now follows from

$$E(y) = (|Q(iy)| + |P(iy)|)(|Q(iy)| - |P(iy)|). \quad \square$$

Examples 3.5. For the implicit midpoint rule, the trapezoidal rule, the Hammer & Hollingsworth, the Kuntzmann & Butcher and Lobatto IIIA methods (c, f, g of Table 3.1) we have $E(y) \equiv 0$ since $Q(z) = P(-z)$. This also follows from Proposition 3.4 because $p = 2j$. A straightforward computation shows that (3.7) is satisfied, hence these methods are A -stable.

For methods d, h, i of Table 3.1 we have $\deg P > \deg Q$ and the leading coefficient of E is negative. Therefore (3.9) cannot be true for $y \rightarrow \infty$ and these methods are not A -stable.

For the Radau IIA method of order 5 (case j) we obtain $E(y) = y^6/3600$ and by inspection of the zeros of $Q(z)$ the method is seen to be A -stable.

For the two-stage SDIRK method (case e) $E(y)$ becomes

$$E(y) = (\gamma - 1/2)^2(4\gamma - 1)y^4. \quad (3.10)$$

Thus the method is A -stable for $\gamma \geq 1/4$. The 3rd order method is A -stable for $\gamma = (3 + \sqrt{3})/6$, but not for $\gamma = (3 - \sqrt{3})/6$ (see Fig. 3.1).

The following general result explains the I -stability properties of the foregoing examples.

Proposition 3.6. *A rational function (3.4) of order $p \geq 2j - 2$ is I -stable if and only if $|R(\infty)| \leq 1$.*

Proof. $|R(\infty)| \leq 1$ implies $k \leq j$. By Proposition 3.4, $E(y)$ must be of the form $K \cdot y^{2j}$. By letting $y \rightarrow \infty$ in (3.6) and (3.9), we see that $|R(\infty)| \leq 1$ is equivalent to $K \geq 0$. \square

L -Stability and $A(\alpha)$ -Stability

The trapezoidal rule for the numerical integration of first-order ordinary differential equations is shown to possess, for a certain type of problem, an undesirable property. (A.R. Gourlay 1970)

A -stability is not the whole answer to the problem of stiff equations. (R. Alexander 1977)

Some of the above methods seem to be optimal in the sense that the stability region coincides *exactly* with the negative half-plane. This property is not as desirable as it may appear, since for a rational function

$$\lim_{z \rightarrow -\infty} R(z) = \lim_{z \rightarrow \infty} R(z) = \lim_{z=iy, y \rightarrow \infty} R(z).$$

The latter must then be 1 in modulus, since $|R(iy)| = 1$ for all real y . This means that for z close to the real axis with a very large negative real part, $|R(z)|$ is, although < 1 , *very close* to one. As a consequence, stiff components in (2.6) are damped out *only very slowly*. We demonstrate this with the example

$$y' = -2000(y - \cos x), \quad y(0) = 0, \quad 0 \leq x \leq 1.5, \quad (3.11)$$

which is the same as (1.1), but with increased stiffness. The numerical results for the trapezoidal rule are compared to those of implicit Euler in Fig. 3.2. The implicit Euler damps out the transient phase much faster than the trapezoidal rule. It thus appears to be a desirable property of a method that $|R(z)|$ be much smaller than 1 for $z \rightarrow -\infty$.

Definition 3.7 (Ehle 1969). A method is called L -stable if it is A -stable and if in addition

$$\lim_{z \rightarrow \infty} R(z) = 0. \quad (3.12)$$

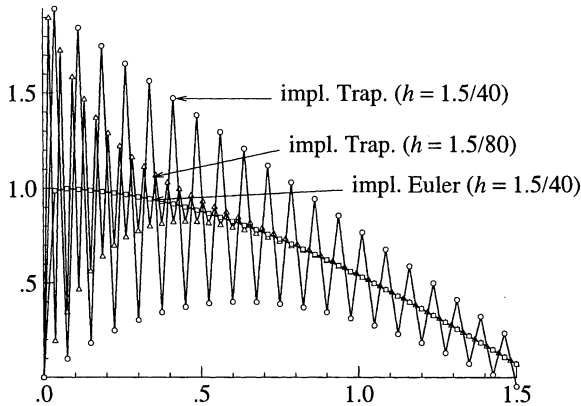


Fig. 3.2. Trapezoidal rule versus implicit Euler on (3.11)

Among the methods of Table 3.1, the implicit Euler, the SDIRK method (e) with $\gamma = (2 \pm \sqrt{2})/2$, as well as the Radau IIA formula (j) are L -stable.

Proposition 3.8. *If an implicit Runge-Kutta method with nonsingular A satisfies one of the following conditions:*

$$a_{sj} = b_j \quad j = 1, \dots, s, \quad (3.13)$$

$$a_{i1} = b_1 \quad i = 1, \dots, s, \quad (3.14)$$

then $R(\infty) = 0$. This makes A -stable methods L -stable.

Proof. By (3.2)

$$R(\infty) = 1 - b^T A^{-1} \mathbb{1} \quad (3.15)$$

and (3.13) means that $A^T e_s = b$ where $e_s = (0, \dots, 0, 1)^T$. Therefore $R(\infty) = 1 - e_s^T \mathbb{1} = 1 - 1 = 0$. In the case of (3.14) use $Ae_1 = \mathbb{1}b_1$. \square

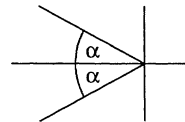
Methods satisfying (3.13) are called *stiffly accurate* (Prothero & Robinson 1974). They are important for the solution of singularly perturbed problems and for differential-algebraic equations (see Chapters VI and VII).

The definition of A -stability is on the one hand too weak, as we have just seen, and on the other hand too strong in the sense that many methods which are not so bad at all are not A -stable. The following definition is a little weaker and will be specially useful in the chapter on multistep methods.

Definition 3.9 (Widlund 1967). A method is said to be $A(\alpha)$ -stable if the sector

$$S_\alpha = \{z; \quad |\arg(-z)| < \alpha, \quad z \neq 0\}$$

is contained in the stability region.



For example, the Padé approximation $R_{03}(z) = \left(1 - z + \frac{z^2}{2!} - \frac{z^3}{3!}\right)^{-1}$ (see (3.29) below) is $A(\alpha)$ -stable for $\alpha \leq 88.23^\circ$.

Numerical Results

To show the effects of good stability properties on the stiff examples of Sect. IV.1, we choose the 3-stage Radau IIA formula (Table 5.6 of Sect. IV.5) which, as we have seen, is A -stable, L -stable and of reasonably high order. It has been coded (Subroutine RADAU5 of the Appendix) and the details of this program will be discussed later (Sect. IV.8). This program integrates all the examples of Sect. IV.1 in a couple of steps and the plots of Fig. 1.3 and Fig. 1.5 show a clear difference.

The beam equation (1.10') with $n = 40$ is integrated, with $Rtol = Atol = 10^{-3}$ (absolute) and smooth initial values, in 28 steps (Fig. 3.3).

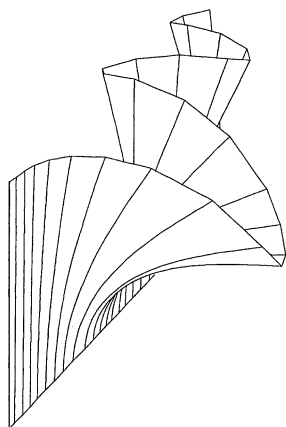


Fig. 3.3. RADAU5 on the beam (1.10'), every step drawn

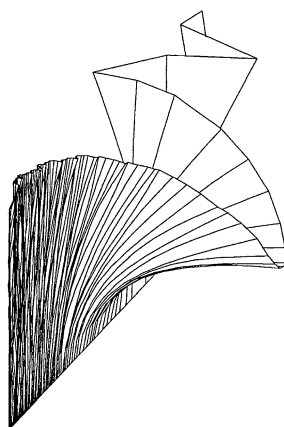


Fig. 3.4. RADAU5 on oscillatory beam with large Tol (107 steps, all drawn)

Since the Radau5 formula is L -stable, the stability domain also covers the imaginary axis and large parts of the right half-plane \mathbb{C}^+ . This means that high oscillations of the true solution *may be damped* by the numerical method. This effect, sometimes judged undesirable (B. Lindberg (1974): “dangerous property . . .”), may also be welcome to suppress uninteresting oscillations. This is demonstrated by applying RADAU5 with very large tolerance ($Rtol = Atol = 1$) to the beam equation (1.10') with $n = 10$ and the perturbed initial value $\theta_n(0) = 0.4$. Here, the high oscillations soon disappear and the numerical solution becomes perfectly smooth (Fig. 3.4). If, however, the tolerance requirement is increased, the program is forced to follow all the oscillations and the picture remains the same as in Fig. 1.11.

Stability Functions of Order $\geq s$

Consider rational functions $R(z) = P(z)/Q(z)$, where $Q(0) = 1$, and both $P(z)$ and $Q(z)$ are polynomials of degree at most s . If $R(z)$ is an approximation of e^z of order $\geq s$, then it follows from (3.5) that

$$e^z Q(z) = P(z) + C_1 z^{s+1} + C_2 z^{s+2} + \dots \quad (3.16)$$

Consequently, the polynomial $P(z)$ and also the error constants C_1, C_2, \dots are uniquely determined in terms of the coefficients of $Q(z)$. For

$$Q(z) = q_0 + q_1 z + q_2 z^2 + \dots + q_s z^s, \quad q_0 = 1 \quad (3.17)$$

an expansion of $e^z Q(z)$ into powers of z yields

$$\begin{aligned} P(z) = q_0 + z \left(\frac{q_0}{1!} + \frac{q_1}{0!} \right) + z^2 \left(\frac{q_0}{2!} + \frac{q_1}{1!} + \frac{q_2}{0!} \right) \\ + \dots + z^s \left(\frac{q_0}{s!} + \frac{q_1}{(s-1)!} + \dots + \frac{q_s}{0!} \right), \end{aligned} \quad (3.18)$$

and for the error constants

$$C_1 = \frac{q_0}{(s+1)!} + \frac{q_1}{s!} + \dots + \frac{q_{s-1}}{2!} + \frac{q_s}{1!} \quad (3.19)$$

$$C_2 = \frac{q_0}{(s+2)!} + \frac{q_1}{(s+1)!} + \dots + \frac{q_{s-1}}{3!} + \frac{q_s}{2!}. \quad (3.20)$$

The Polynomial $M(x)$. With help of the polynomial

$$M(x) = q_s + q_{s-1} \frac{x}{1!} + q_{s-2} \frac{x^2}{2!} + \dots + q_0 \frac{x^s}{s!} \quad (3.21)$$

the formulas for $Q(z)$ and $P(z)$ become more symmetric. We have

$$Q(z) = M^{(s)}(0) + M^{(s-1)}(0)z + \dots + M(0)z^s \quad (3.22)$$

$$P(z) = M^{(s)}(1) + M^{(s-1)}(1)z + \dots + M(1)z^s, \quad (3.23)$$

and the error constants are given by

$$C_1 = \int_0^1 M(x) dx, \quad C_2 = \int_0^1 (1-x)M(x) dx. \quad (3.24)$$

For the stability function of collocation methods we have the following nice result.

Theorem 3.10 (K. Wright 1970, S.P. Nørsett 1975). *The stability function of the collocation method based on the points c_1, c_2, \dots, c_s is given by $R(z) = P(z)/Q(z)$, where $Q(z)$ and $P(z)$ are the polynomials of (3.22) and (3.23), respectively, with $M(x)$ given by*

$$M(x) = \frac{1}{s!} \prod_{i=1}^s (x - c_i). \quad (3.25)$$

Proof (Nørsett & Wanner 1979). We assume $x_0 = 0$, $h = 1$, $\lambda = z$, $y_0 = 1$ and let $u(x)$ be the collocation polynomial. Since $u'(x) - zu(x)$ is a polynomial of degree s which vanishes at the collocation points, there are constants K_0 and K such that

$$u'(x) - zu(x) = K_0 M(x) \quad \text{or} \quad \left(1 - \frac{D}{z}\right)u(x) = K M(x) \quad (3.26)$$

with the polynomial $M(x)$ of (3.25) (D denotes the differentiation operator). Expanding $(1 - D/z)^{-1}$ into a geometric series yields

$$u(x) = K \left(1 + \frac{D}{z} + \frac{D^2}{z^2} + \dots + \frac{D^s}{z^s}\right) M(x), \quad (3.27)$$

because $M^{(j)}(x) \equiv 0$ for $j > s$. From $u(1) = R(z)u(0)$ we have the relation $R(z) = u(1)/u(0)$, which leads to (3.22) and (3.23). \square

Padé Approximations to the Exponential Function

Comme cela est souvent le cas en ce qui concerne les découvertes scientifiques, leur inventeur n'est pas H. Padé.

(C. Brezinski 1984, Œuvres de H. Padé, p. 5)

Padé approximations (Padé 1892) are rational functions which, for a given degree of the numerator and the denominator, have highest order of approximation. Their origin lies in the theory of continued fractions and they played a fundamental role in Hermite's (1873) proof of the transcendence of e .

These optimal approximations can be obtained for the exponential function e^z from (3.22) and (3.23) by the following idea (Padé 1899): choose $M(x)$ such that in (3.22) and (3.23) as many terms as possible involving high powers of z become zero, i.e.,

$$M(x) = \frac{x^k(x-1)^j}{(k+j)!}; \quad (3.28)$$

then $M^{(i)}(0) = 0$ for $i = 0, \dots, k-1$ and $M^{(i)}(1) = 0$ for $i = 0, \dots, j-1$.

Theorem 3.11. *The (k, j) -Padé approximation to e^z is given by*

$$R_{kj}(z) = \frac{P_{kj}(z)}{Q_{kj}(z)} \quad (3.29)$$

where

$$\begin{aligned} P_{kj}(z) &= 1 + \frac{k}{j+k}z + \frac{k(k-1)}{(j+k)(j+k-1)} \cdot \frac{z^2}{2!} + \dots + \frac{k(k-1)\dots 1}{(j+k)\dots(j+1)} \cdot \frac{z^k}{k!} \\ Q_{kj}(z) &= 1 - \frac{j}{k+j}z + \frac{j(j-1)}{(k+j)(k+j-1)} \cdot \frac{z^2}{2!} - \dots + (-1)^j \frac{j(j-1)\dots 1}{(k+j)\dots(k+1)} \cdot \frac{z^j}{j!} \\ &= P_{jk}(-z), \end{aligned}$$

with error

$$e^z - R_{kj}(z) = (-1)^j \frac{j!k!}{(j+k)!(j+k+1)!} z^{j+k+1} + \mathcal{O}(z^{j+k+2}). \quad (3.30)$$

It is the unique rational approximation to e^z of order $j+k$, such that the degrees of numerator and denominator are k and j , respectively.

Table 3.2. Padé approximations for e^z

$\frac{1}{1}$	$\frac{1+z}{1}$	$\frac{1+z+\frac{z^2}{2!}}{1}$
$\frac{1}{1-z}$	$\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$	$\frac{1+\frac{2}{3}z+\frac{1}{3}\frac{z^2}{2!}}{1-\frac{1}{3}z}$
$\frac{1}{1-z+\frac{z^2}{2!}}$	$\frac{1+\frac{1}{3}z}{1-\frac{2}{3}z+\frac{1}{3}\frac{z^2}{2!}}$	$\frac{1+\frac{1}{2}z+\frac{1}{6}\frac{z^2}{2!}}{1-\frac{1}{2}z+\frac{1}{6}\frac{z^2}{2!}}$
$\frac{1}{1-z+\frac{z^2}{2!}-\frac{z^3}{3!}}$	$\frac{1+\frac{1}{4}z}{1-\frac{3}{4}z+\frac{1}{2}\frac{z^2}{2!}-\frac{1}{4}\frac{z^3}{3!}}$	$\frac{1+\frac{2}{5}z+\frac{1}{10}\frac{z^2}{2!}}{1-\frac{3}{5}z+\frac{3}{10}\frac{z^2}{2!}-\frac{1}{10}\frac{z^3}{3!}}$

Proof. Inserting (3.28) into (3.22) and (3.23) gives the formulas for $P_{kj}(z)$, $Q_{kj}(z)$ and (3.30). The uniqueness is a consequence of the fact that the $(j+k)$ -degree polynomial $M(x)$ of (3.21) must have a zero of multiplicity k at $x=0$, and one of multiplicity j at $x=1$. \square

Table 3.2 shows the first Padé approximations to e^z . We observe that the stability function of many methods of Table 3.1 are Padé approximations. The *diagonal Padé approximations* are those with $k=j$.

Exercises

1. Let $R(z)$ be the stability function of (3.1) and $R^*(z)$ the stability function of its adjoint method (see Sect. II.8). Prove that

$$R^*(z) = (R(-z))^{-1}.$$

2. Consider an implicit Runge-Kutta method of order $p \geq s$ with nonsingular A , distinct c_i and non-zero b_i . Show
 - a) If $C(s)$ and $c_s = 1$ then (3.13);
 - b) If $D(s)$ and $c_1 = 0$ then (3.14).

In both cases the stability function satisfies $R(\infty) = 0$.

(For the definition of the assumptions $C(s)$ and $D(s)$ see Sect. IV.5).

3. Show that collocation methods can only be L -stable if $M(1) = 0$, i.e., if one of the c 's, usually c_s , equals 1.
4. (Padé (1899), see also Lagrange (1776)). Show that the continued fraction

$$e^x = 1 + \frac{x}{1 - \frac{x}{2} + \frac{\frac{1}{1 \cdot 3} \frac{x^2}{4}}{1 + \frac{\frac{1}{3 \cdot 5} \frac{x^2}{4}}{1 + \frac{\frac{1}{5 \cdot 7} \frac{x^2}{4}}{1 + \frac{\frac{1}{7 \cdot 9} \frac{x^2}{4}}{1 + \dots}}}}}$$

leads to the diagonal Padé approximations for e^x .

Hint. Compute the first partial fractions. If you don't succeed in finding a general proof, read Sect. IV.5.

5. The trapezoidal rule

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

satisfies $a_{si} = b_i$, but not $R(\infty) = 0$. Why doesn't this contradict Proposition 3.8?

6. Show that

$$\begin{aligned} y_1 &= y_0 + hf(y_0 + \theta(y_1 - y_0)) \\ y_1 &= y_0 + h(1 - \theta)f(y_0) + h\theta f(y_1) \end{aligned}$$

are both nonlinear extensions of the θ -method. Find others.

7. The composition of a step of the θ -method with step-size αh , followed by a θ' -method with step-size $(1 - 2\alpha)h$ and again a θ -method with step-size αh leads to

$$R(z) = \left(\frac{1 + \alpha z(1 - \theta)}{1 - \alpha z\theta} \right)^2 \left(\frac{1 + (1 - 2\alpha)z(1 - \theta')}{1 - (1 - 2\alpha)z\theta'} \right)$$

Show that this method, for $\theta' = 1 - \theta$, is of order 2 if $\alpha = 1 - \sqrt{2}/2$ and strongly A -stable (i.e., A -stable and $|R(\infty)| < 1$) for $\theta > 1/2$. The authors Müller, Prohl, Rannacher & Turek (1994) call this method "fractional θ -method" and use it successfully for computations of the incompressible Navier-Stokes equations.

IV.4 Order Stars

Mein hochgeehrter Lehrer, der vor wenigen Jahren verstorbene Geheime Hofrath *Gauss* in Göttingen, pflegte in vertraulichem Gespräche häufig zu äussern, die Mathematik sei weit mehr eine Wissenschaft für das Auge als eine für das Ohr. Was das Auge mit einem Blicke sogleich übersieht . . .

(J.F. Encke 1861, publ. in Kronecker's Werke, Vol. 5, page 391)

Order stars, discovered by searching for a better understanding of the stability properties of the Padé approximations to e^z (Wanner, Hairer & Nørsett 1978), offered nice and unexpected access to many other results: the “second barrier” of Dahlquist, the Daniel & Moore conjecture, highest possible order with real poles, comparison of stability domains (Jeltsch & Nevanlinna 1981, 1982), order bounds for hyperbolic or parabolic difference schemes (e.g., Iserles & Strang 1983, Iserles & Williamson 1983, Jeltsch 1988).

Introduction

When I wrote my book in 1971 I wanted to draw “relative stability domains”, but curious stars came out from the plotter. I thought of an error in the program and I threw them away . . .

(C.W. Gear, in 1979)

We present in Fig. 4.1 the stability domains for the Padé approximations R_{33} , R_{24} , R_{15} , R_{06} of Theorem 3.12, which are all 6th order approximations to $\exp(z)$. It can be observed that R_{33} and R_{24} are nicely A -stable. The other two are not, R_{15} violates (3.6) and R_{06} violates (3.7). After some meditation on these and similar figures, trying to obtain a better understanding of these phenomena, one is finally led to

Definition 4.1. The set

$$A = \left\{ z \in \mathbb{C} ; |R(z)| > |e^z| \right\} = \left\{ z \in \mathbb{C} ; |q(z)| > 1 \right\} \quad (4.1)$$

where $q(z) = R(z)/e^z$, is called the *order star* of R .

The order star does not compare $|R(z)|$ to 1, as does the stability domain, but to the exact solution $|e^z| = e^x$ and it is hoped that this might give more information. As we always assume that the coefficients of $R(z)$ are real, the order star is symmetric with respect to the real axis. Furthermore, since $|e^{iy}| = 1$, A is the complementary set of the stability domain S on the imaginary axis. Therefore we have from (3.6) and (3.7):

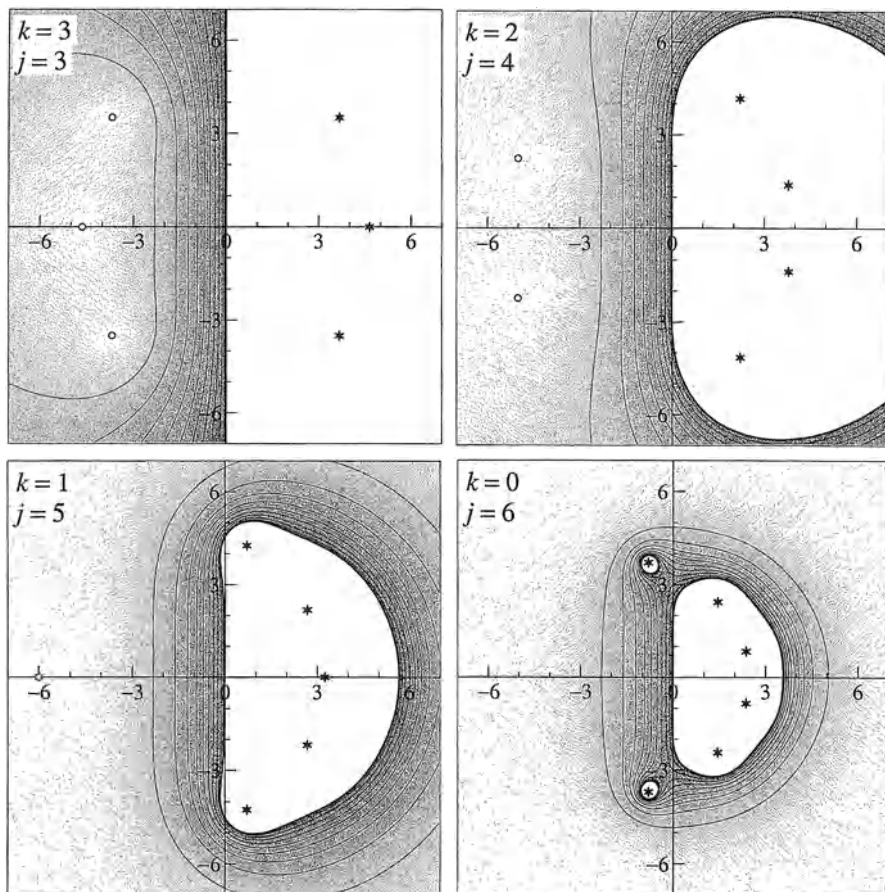


Fig. 4.1. Stability domains for Padé approximations

Lemma 4.2. $R(z)$ is I -stable if and only if

- (i) $A \cap i\mathbb{R} = \emptyset$.

Further, $R(z)$ is A -stable if and only if (i) and

- (ii) all poles of $R(z)$ (= poles of $q(z)$) lie in the positive half plane \mathbb{C}^+ . \square

Fig. 4.2 shows the order stars corresponding to the functions of Fig. 4.1. These order stars show a nice and regular behaviour: there are j black “fingers” to the right, each containing a pole of R_{kj} , and k white “fingers” to the left, each containing a zero. Exactly two boundary curves of A tend to infinity near to the imaginary axis. These properties are a consequence of the following three Lemmas.

Lemma 4.3. If $R(z)$ is an approximation to e^z of order p , i.e., if

$$e^z - R(z) = Cz^{p+1} + \mathcal{O}(z^{p+2}) \quad (4.2)$$

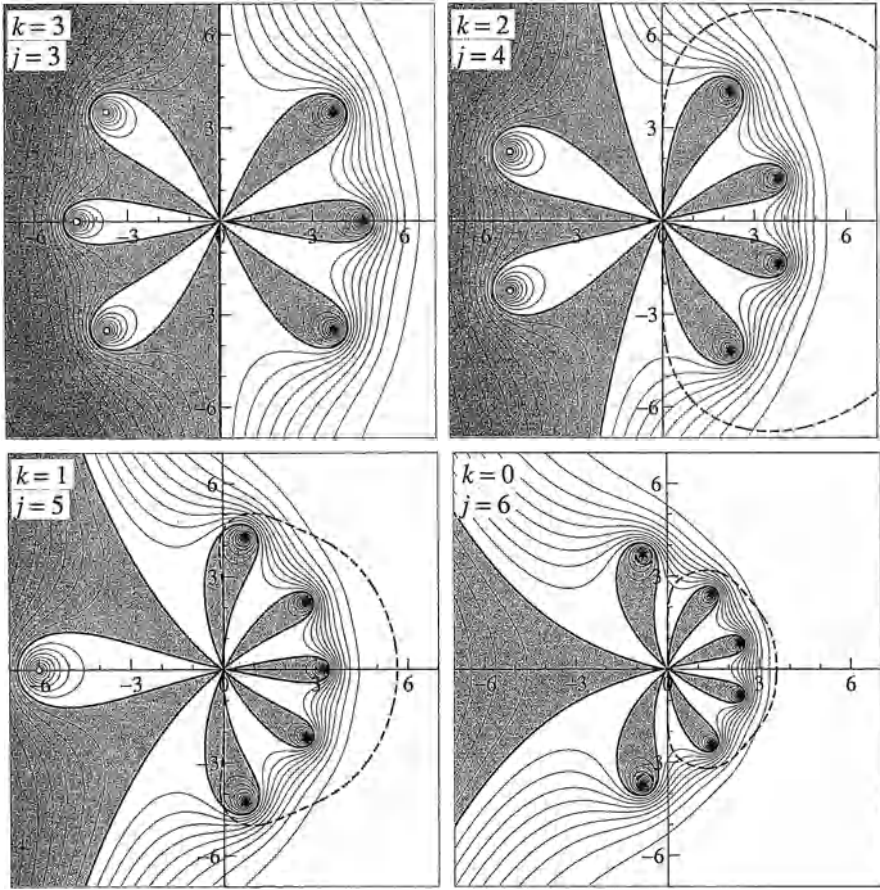


Fig. 4.2. Order stars for Padé approximations

with $C \neq 0$, then, for $z \rightarrow 0$, A behaves like a “star” with $p+1$ sectors of equal width $\pi/(p+1)$, separated by $p+1$ similar “white” sectors of the complementary set. The positive real axis is inside a black sector iff $C < 0$ and inside a white sector iff $C > 0$.

Proof. Dividing the error formula (4.2) by e^z gives

$$\frac{R(z)}{e^z} = 1 - Cz^{p+1} + \mathcal{O}(z^{p+2}).$$

Thus the value $R(z)/e^z$ surrounds the point 1 as often as z^{p+1} surrounds the origin, namely $p+1$ times. So, $R(z)/e^z$ is $p+1$ times alternatively inside or outside the unit circle. It lies inside for small positive real z whenever $C > 0$. \square

Lemma 4.4. *If $z = re^{i\theta}$ and $r \rightarrow \infty$, then $z \in A$ for $\pi/2 < \theta < 3\pi/2$ and $z \notin A$ for $-\pi/2 < \theta < \pi/2$. The border ∂A possesses only two branches which go to infinity. If*

$$R(z) = Kz^\ell + \mathcal{O}(z^{\ell-1}) \quad \text{for } z \rightarrow \infty, \quad (4.3)$$

these branches asymptotically approach

$$x = \log |K| + \ell \log |y| \quad (4.4)$$

Proof. The first assertion is the well-known fact that the exponential function, for $\operatorname{Re} z \rightarrow \pm\infty$ is much stronger than any polynomial or rational function. In order to show the uniqueness of the border lines, we consider for $r \rightarrow \infty$ the two functions

$$\begin{aligned} \varphi_1(\theta) &= |e^z|^2 = e^{2r \cos \theta} \\ \varphi_2(\theta) &= |R(z)|^2 = R(re^{i\theta})R(re^{-i\theta}). \end{aligned}$$

Differentiation gives

$$\frac{\varphi'_1}{\varphi_1} = -2r \sin \theta, \quad \frac{\varphi'_2}{\varphi_2} = 2r \operatorname{Re} \left(ie^{i\theta} \cdot \frac{R'(re^{i\theta})}{R(re^{i\theta})} \right). \quad (4.5)$$

Since $|R'/R| \rightarrow 0$ for $r \rightarrow \infty$, we have

$$\frac{d}{d\theta} \log \varphi_1(\theta) < \frac{d}{d\theta} \log \varphi_2(\theta) \quad \text{for } \theta \in [\varepsilon, \pi - \varepsilon].$$

Hence in this interval there can only be one value of θ with $\varphi_1(\theta) = \varphi_2(\theta)$. Formula (4.4) is obtained from (4.3) by

$$|K|(x^2 + y^2)^{\ell/2} \approx e^x, \quad \log |K| + \frac{\ell}{2} \log(x^2 + y^2) \approx x$$

and by neglecting x^2 , which is justified because $x/y \rightarrow 0$ whenever $x + iy$ tends to infinity on the border of A . \square

It is clear from the maximum principle that each bounded “finger” of A in Fig. 4.2 must contain a pole of $q(z)$. A still stronger result is the following:

Lemma 4.5. *Each bounded subset $F \subset A$ with common boundary $\partial F \subset \partial A$ collecting m sectors at the origin must contain at least m poles of $q(z)$ (each counted according to its multiplicity). Analogously, each bounded “white” subset $F \subset \mathbb{C} \setminus A$ with m sectors at the origin must contain at least m zeros of $q(z)$.*

Proof. Suppose first that ∂F is represented by a parametrized positively oriented loop $c(t)$, $t_0 \leq t \leq t_1$. Let $\vec{a} = (c'_1(t), c'_2(t))$ be the tangent vector and $\vec{n} = (c'_2(t), -c'_1(t))$ an exterior normal vector. We write

$$q(z) = r(x, y) \cdot e^{i\varphi(x, y)}, \quad z = x + iy$$

so that $\log q(z) = \log r(x, y) + i\varphi(x, y)$. Since the modulus increases inside F , we have

$$\frac{\partial(\log r)}{\partial \vec{n}} \leq 0. \quad (4.6)$$

Now the Cauchy-Riemann differential equations for $\log q$ are

$$\frac{\partial(\log r)}{\partial x} = \frac{\partial \varphi}{\partial y}; \quad \frac{\partial(\log r)}{\partial y} = -\frac{\partial \varphi}{\partial x}, \quad (4.7)$$

so that (4.6) becomes

$$\frac{\partial \varphi}{\partial \vec{n}} \leq 0. \quad (4.8)$$

This inequality is strict except at a finite number of points, because $q'(c(t)) \cdot c'(t) = i \cdot q(c(t)) \cdot \partial \varphi / \partial \vec{a}$ and the number of zeros of $q'(z)$ is finite. Thus the *argument* of q decreases along c . If the contour curve $c(t)$ returns m times to the origin, where the argument is a multiple of 2π , the vector $q(z)$ must perform at least m complete revolutions in the negative sense (Fig. 4.3). Thus the argument principle (an idea which we have already encountered in Sect. I.13; see Volume I, pages 81 and 382), ensures the presence of at least m poles inside F (there are no zeros, because these are not in A).

If the boundary curve is represented by several curves, all rotation numbers are added up. For “white” subsets the proof is similar, just that $\partial(\log r) / \partial \vec{n} > 0$ and the argument rotates in the other sense. \square

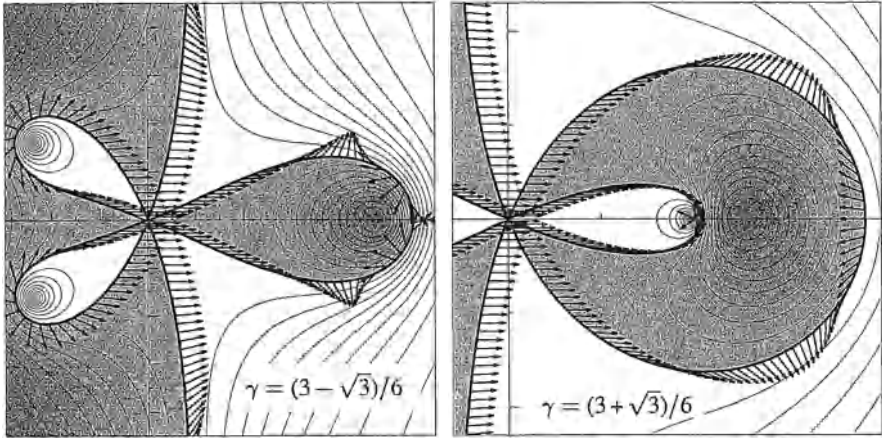


Fig. 4.3. SDIRK methods, order 3; arrows indicate direction of $q(z)$

Fig. 4.3 gives an illustration of two order stars for the SDIRK methods of order 3 (Table 3.1, case e). Here, $q(z)$ possesses a double pole at $z = 1/\gamma$. However, for $\gamma = (3 - \sqrt{3})/6$, the bounded component F of A collects only *one* sector at the origin. Since the vector $q(z)$ performs two rotations, there is in addition to

the origin a second point on ∂F for which $\arg(q) = 0$, i.e., $\arg(R(z)) = \arg(e^z)$. Thus, because $|R(z)| = |e^z|$ on ∂A , we have $R(z) = e^z$. These points are called *exponential fitting points*. Another version of Lemma 4.5 is thus (Iserles 1981):

Lemma 4.5'. *Each bounded subset $F \subset A$ with $\partial F \subset \partial A$ contains exactly as many poles as there are exponential fitting points on its boundary.* \square

Order and Stability for Rational Approximations

In the sequel we suppose $R(z)$ to be an arbitrary rational approximation of order p with k zeros and j poles.

Theorem 4.6. *If $R(z)$ is A -stable, then $p \leq 2k_1 + 2$, where k_1 is the number of different zeros of $R(z)$ in \mathbb{C}^- .*

Proof. At least $[(p+1)/2]$ sectors of A start in \mathbb{C}^- (Lemma 4.3). By A -stability these have to be infinite and enclose at least $[(p+1)/2] - 1$ bounded white fingers, each containing at least one zero by Lemma 4.5. Therefore $[(p+1)/2] - 1 \leq k_1$. \square

Theorem 4.7. *If $R(z)$ is I -stable, then $p \leq 2j_1$, where j_1 is the number of poles of $R(z)$ in \mathbb{C}^+ .*

Proof. At least $[(p+1)/2]$ sectors of A start in \mathbb{C}^+ . They cannot cross $i\mathbb{R}$ and must therefore be bounded (Lemma 4.4). Again by Lemma 4.5 we have $[(p+1)/2] \leq j_1$. \square

Theorem 4.8. *Suppose that $p \geq 2j - 1$ and $|R(\infty)| \leq 1$. Then, $R(z)$ is A -stable.*

Proof. By Proposition 3.6 the function $R(z)$ is I -stable. Applying Theorem 4.7 we get $j_1 \geq j$ so that I -stability implies A -stability. \square

Theorem 4.9 (Crouzeix & Ruamps 1977). *Suppose $p \geq 2j - 2$, $|R(\infty)| \leq 1$, and the coefficients of the denominator $Q(z)$ have alternating signs. Then, $R(z)$ is A -stable.*

Proof. A similar argument as in the foregoing proof allows at most one pole in \mathbb{C}^- . It would then be real and its existence would contradict the hypothesis on signs of $Q(z)$. \square

Theorem 4.10. Suppose $p \geq 2j - 3$, $R(z)$ is I -stable, and the coefficients of $Q(z)$ have alternating signs. Then, $R(z)$ is A -stable.

Proof. For $p \geq 2j - 3$ the argument of the foregoing proof is still valid. However Proposition 3.6 is no longer applicable and we need the hypothesis on I -stability. \square

We see from Fig. 4.2 that all poles and all zeros for Padé approximations must be *simple*. Whenever two poles coalesce, the corresponding sectors create a bounded white finger between them with the need for an additional zero. Thus the presence of multiple zeros or poles will require an order reduction.

Theorem 4.11. Let $R(z)$ possess k_0 distinct zeros and j_0 distinct poles. Then, $p \leq k_0 + j_0$.

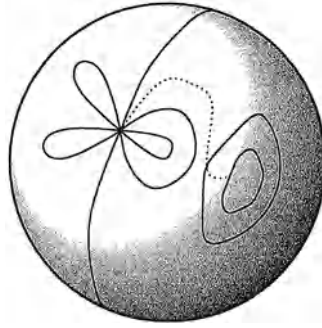


Fig. 4.4. Order star on Gaussian sphere

Proof. We identify the complex plane with the Gaussian sphere and the order star with a CW-complex decomposition of this sphere (Fig. 4.4). Let s_2 be the number of 2-cells f_i , s_1 the number of 1-cells l_i (paths), and s_0 the number of vertices. Then Euler's polyhedral formula ("Si enim numerus angularum solidorum fuerit = S , numerus acierum = A et numerus hedrarum = H , semper habetur $S + H = A + 2$, hincque vel $S = A + 2 - H$ vel $H = A + 2 - S$ vel $A = S + H - 2$, quae relationis simplicitas ob demonstrationis difficultatem . . .", Euler (1752)), implies

$$s_0 - s_1 + s_2 = 2. \quad (4.9)$$

Modern versions are in any book on algebraic topology, for particularly easy reading see e.g. Massey (1980, p. 87, Corollary 4.4). Formula (4.9) is only true if all f_i are homeomorphic to disks. Otherwise, they have to be cut into disks by additional paths (dotted in Fig. 4.4). So, in general, we have

$$s_0 - s_1 + s_2 \geq 2. \quad (4.9')$$

Since each vertex is reached by at least 2 paths, the origin by hypothesis by $2p + 2$, and since every path has two extremities, we have

$$s_1 - s_0 \geq p. \quad (4.10)$$

By Lemma 4.5 each 2-cell, with the exception of two (the two “infinite” ones) must contain at least a pole or a zero, so we have

$$s_2 \leq k_0 + j_0 + 2. \quad (4.11)$$

These three inequalities give $p \leq k_0 + j_0$. □

Stability of Padé Approximations

... evidence is given to suggest that these are the only L-acceptable Padé approximations to the exponential.

(B.L. Ehle 1973)

Theorem 4.12. *A Padé approximation $R_{k,j}(z)$, given in (3.30), is A -stable if and only if $k \leq j \leq k + 2$. All zeros and all poles are simple.*

Proof. The “if”-part is a consequence of Theorem 4.9. The “only if”-part follows from Theorem 4.6 since $p = k + j$. For the same reason Theorem 4.11 shows that all poles and zeros are simple. □

Comparing Stability Domains

Da ist der allerärmste Mann
dem ander'n viel zu reich,
das Schicksal setzt den Hobel an
und hobelt beide gleich.

(F. Raimund, das Hobellied)

Jeltsch & Nevanlinna (1978) proved the following “disk theorem”: *If S is the stability domain of an s -stage explicit Runge-Kutta method and D the disk with centre $-s$ and radius s (i.e the stability domain of s explicit Euler steps with step size h/s), then*

$$S \not\supset D \quad (4.12)$$

unless $S = D$ and the method in question is Euler's method. This curious result expresses the fact that Euler's method is “the most stable” of all methods with equal numerical work. After the discovery of order stars it became clear that the result is much more general and that *any* method has the same property (Jeltsch & Nevanlinna 1981). We shall also see in Chapter V that this result generalizes to many multistep methods. The main tool of this theory is

Definition 4.13. Let $R_1(z)$ and $R_2(z)$ be rational approximations to e^z , then their *relative order star* is defined as

$$B = \left\{ z \in \mathbb{C} ; \left| \frac{R_1(z)}{R_2(z)} \right| > 1 \right\}. \quad (4.13)$$

Here, the stability function for method 1 is compared to the stability function for method 2 instead of to the exact solution e^z . The following order relations

$$e^z - R_1(z) = C_1 z^{p_1+1} + \dots$$

$$e^z - R_2(z) = C_2 z^{p_2+1} + \dots$$

lead, by subtraction, to

$$\frac{R_1(z)}{R_2(z)} = 1 - C z^{p+1} + \dots \quad (4.14)$$

where $p = \min(p_1, p_2)$ and

$$C = \begin{cases} C_1 - C_2 & \text{if } p_1 = p_2 \\ C_1 & \text{if } p_1 < p_2 \\ -C_2 & \text{if } p_1 > p_2. \end{cases} \quad (4.15)$$

Remark 4.14. The statement of Lemma 4.3 remains unchanged for B , whenever $C \neq 0$. Since the fraction $R_1(z)/R_2(z)$ has no essential singularity at infinity, there is no analogue of Lemma 4.4. Further, the boundedness assumption on F can be omitted in Lemmas 4.5 and 4.5' (if ∞ is a pole of $R_1(z)/R_2(z)$, it has to be counted also). With the correspondences displayed in Table 4.1, the statements of Theorems 4.6 and 4.7 remain true for B .

Table 4.1. Correspondences between A and B

order star A (4.1)	\longleftrightarrow	relative order star B (4.13)
imaginary axis	\longleftrightarrow	∂S_2
\mathbb{C}^-	\longleftrightarrow	interior of S_2
\mathbb{C}^+	\longleftrightarrow	exterior of S_2
method A-stable	\longleftrightarrow	$S_1 \supset S_2$
p	\longleftrightarrow	$\min(p_1, p_2)$

Theorem 4.15. If $R_1(z)$ and $R_2(z)$ are polynomial stability functions of degree s and orders ≥ 1 , then the corresponding stability domains satisfy

$$S_1 \not\supset S_2 \quad \text{and} \quad S_1 \not\subset S_2. \quad (4.16)$$

Proof. Suppose that $S_1 \supset S_2$ (i.e., by Table 4.1, suppose “A-stability”). Then the analogue of Theorem 4.7 requires that $R_1(z)/R_2(z)$ have a pole *outside* S_2 . Since

$R_1(z)$ and $R_2(z)$ have the same degree, $R_1(z)/R_2(z)$ has no pole at infinity. Therefore the only poles of $R_1(z)/R_2(z)$ are the zeros of $R_2(z)$ and these are *inside* S_2 . This is a contradiction and proves the first part of (4.16). The second part is obtained by exchanging $R_1(z)$ and $R_2(z)$. \square

In order to compare numerical methods with *different* numerical work, we consider scaled stability domains.

Definition 4.16. Let $R(z)$ be the stability function of degree s of an explicit Runge-Kutta method (usually with s stages), then

$$S^{scal} = \left\{ z ; |R(sz)| \leq 1 \right\} = \left\{ z ; s \cdot z \in S \right\} = \frac{1}{s} S \quad (4.17)$$

will be called the *scaled stability domain* of the method.

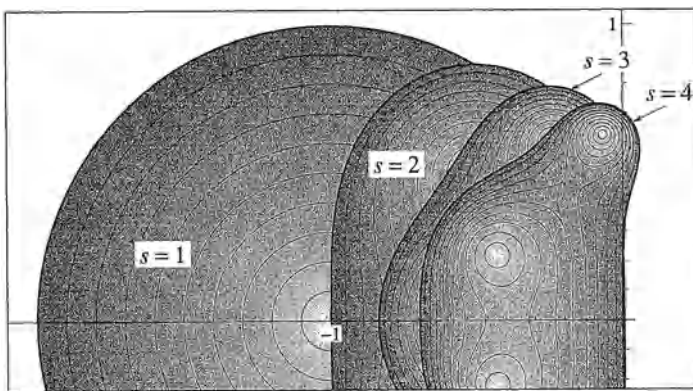


Fig. 4.5. Scaled stability domains for Taylor methods (2.12)

Theorem 4.17 (Jeltsch & Nevanlinna 1981). If $R_1(z)$ and $R_2(z)$ are the stability functions of degrees s_1 resp. s_2 of two explicit Runge-Kutta methods of orders ≥ 1 , then

$$S_1^{scal} \not\supset S_2^{scal} \quad \text{and} \quad S_1^{scal} \not\subset S_2^{scal}, \quad (4.18)$$

i.e., a scaled stability domain can never completely contain another.

The interesting interpretation of this result is that for any two methods, there exists a differential equation $y' = \lambda y$ such that one of them performs better than the other. No “miracle” method is possible.

Proof. We compare s_2 steps of method 1 with step size h/s_2 to s_1 steps of method 2 with step size h/s_1 . Both procedures then have comparable numerical work for the same advance in step size. Applied to $y' = \lambda y$, this compares

$$\left(R_1\left(\frac{z}{s_2}\right) \right)^{s_2} \quad \text{to} \quad \left(R_2\left(\frac{z}{s_1}\right) \right)^{s_1}$$

of the same degree. Theorem 4.15 now gives

$$s_2 \cdot S_1 \not\preceq s_1 \cdot S_2 \quad \text{or} \quad S_1^{scal} \not\preceq S_2^{scal}. \quad \square$$

As an illustration to this theorem, we present in Fig. 4.5 the *scaled* stability domains for the Taylor methods of orders 1, 2, 3, 4 (compare with Fig. 2.1). It can clearly be observed that none of them contains another.

Rational Approximations with Real Poles

The surprising result is that the maximum reachable order is $m + 1$.
(Nørsett & Wolfbrandt 1977)

The stability functions of diagonally implicit Runge-Kutta methods (DIRK methods), i.e., methods with $a_{ij} = 0$ for $i < j$, are

$$R(z) = \frac{P(z)}{(1 - \gamma_1 z)(1 - \gamma_2 z) \dots (1 - \gamma_s z)}, \quad (4.19)$$

where $\gamma_i = a_{ii}$ ($i = 1, \dots, s$) and degree $P \leq s$. This follows at once from Formula (3.3) of Proposition 3.2, since the determinant of a triangular matrix is the product of its diagonal elements. Thus $R(z)$ possesses *real poles* $1/\gamma_1, 1/\gamma_2, \dots, 1/\gamma_s$. Such approximations to e^z will also appear in the next sections as stability functions of Rosenbrock methods and so-called singly-implicit Runge-Kutta methods. They thus merit a more thorough study. Research on these real-pole approximations was started by Nørsett (1974) and Wolfbrandt (1977). Many results are collected in their joint paper Nørsett & Wolfbrandt (1977).

If the method is of order at least s , $P(z)$ is given by (3.18). We shall here, and in the sequel, very often write the formulas for $s = 3$ without always mentioning how trivial their extension to arbitrary s is. Hence for $s = 3$

$$R(z) = \frac{1 + z\left(\frac{S_0}{1!} - \frac{S_1}{0!}\right) + z^2\left(\frac{S_0}{2!} - \frac{S_1}{1!} + \frac{S_2}{0!}\right) + z^3\left(\frac{S_0}{3!} - \frac{S_1}{2!} + \frac{S_2}{1!} - \frac{S_3}{0!}\right)}{1 - zS_1 + z^2S_2 - z^3S_3} \quad (4.20)$$

where

$$S_0 = 1, \quad S_1 = \gamma_1 + \gamma_2 + \gamma_3, \quad S_2 = \gamma_1\gamma_2 + \gamma_1\gamma_3 + \gamma_2\gamma_3, \quad S_3 = \gamma_1\gamma_2\gamma_3.$$

The error constant is for $p = s$

$$C = \frac{S_0}{4!} - \frac{S_1}{3!} + \frac{S_2}{2!} - \frac{S_3}{1!}. \quad (4.21)$$

Theorem 4.18. *Let $R(z)$ be an approximation to e^z of order p with real poles only and let k be the degree of its numerator. Then,*

$$p \leq k + 1.$$

Proof. If a sector of the order star A ends up with a pole on the real axis, then by symmetry the complex conjugate sector must join the first one. All white sectors enclosed by these two must therefore be finite (Fig. 4.6.). The same is true for sectors joining the infinite part of A . There is thus on each side of the real axis at most one white sector which can be infinite. Thus the remaining $p - 1$ white sectors require together at least $p - 1$ zeros by Lemma 4.5, i.e., we have $p - 1 \leq k$. \square

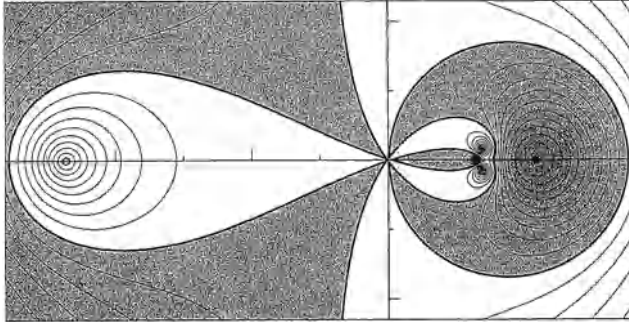


Fig. 4.6. An approximation with real poles, 3 zeros, order 4

Remark 4.19. If $p \geq k$, then at least one white sector must be unbounded. This is then either the first sector on the positive real axis, or, by symmetry, there is a pair of two sectors. By the proof of Theorem 4.18 the pair is unique and we shall call it *Cary Grant's part*.



Remark 4.20. If $p = k + 1$, the optimal case, there are $k + 2$ white sectors, two of them are infinite. Hence each of the remaining k sectors must then contain exactly one root of $P(z)$. As a consequence, $C < 0$ iff $P(z)$ has no positive real root between the origin and the first pole.

The Real-Pole Sandwich

We now analyze the approximations (4.19) with order $p \geq s$ in more detail (Nørsett & Wanner 1979). We are interested in two sets:

Definition 4.21. Let L be the set of $(\gamma_1, \dots, \gamma_s)$ for which $\deg P(z)$ in (4.20) is $\leq s - 1$, i.e., $R(\infty) = 0$ for $\gamma_i \neq 0$ ($i = 1, \dots, s$).

Definition 4.22. Denote by H the set of $(\gamma_1, \dots, \gamma_s)$ for which the error constant (4.21) is zero, i.e., for which the approximation has highest possible order $p = s + 1$.

A consequence of Theorem 4.18 is

$$L \cap H = \emptyset. \quad (4.22)$$

Written for the case $s = 3$ (generalizations to arbitrary s are straightforward) and using (4.20) and (4.21) the sets L and H become

$$\begin{aligned} L &= \left\{ (\gamma_1, \gamma_2, \gamma_3); \frac{1}{3!} - \frac{\gamma_1 + \gamma_2 + \gamma_3}{2!} + \frac{\gamma_1 \gamma_2 + \gamma_1 \gamma_3 + \gamma_2 \gamma_3}{1!} - \frac{\gamma_1 \gamma_2 \gamma_3}{0!} = 0 \right\} \\ H &= \left\{ (\gamma_1, \gamma_2, \gamma_3); \frac{1}{4!} - \frac{\gamma_1 + \gamma_2 + \gamma_3}{3!} + \frac{\gamma_1 \gamma_2 + \gamma_1 \gamma_3 + \gamma_2 \gamma_3}{2!} - \frac{\gamma_1 \gamma_2 \gamma_3}{1!} = 0 \right\}. \end{aligned} \quad (4.23)$$

Theorem 4.23 (Nørsett & Wanner 1979). *The surfaces H and L are each composed of s disjoint connected sheets*

$$L = L_1 \cup L_2 \cup \dots \cup L_s, \quad H = H_1 \cup H_2 \cup \dots \cup H_s. \quad (4.24)$$

If a direction $\delta = (\delta_1, \dots, \delta_s)$ is chosen with all $\delta_i \neq 0$ and if k of them are positive, then the ray

$$X = \left\{ (\gamma_1, \dots, \gamma_s); \gamma_i = t\delta_i, \quad 0 \leq t < \infty \right\} \quad (4.25)$$

intersects the sheets $H_1, L_1, H_2, L_2, \dots, H_k, L_k$ in this order and no others.

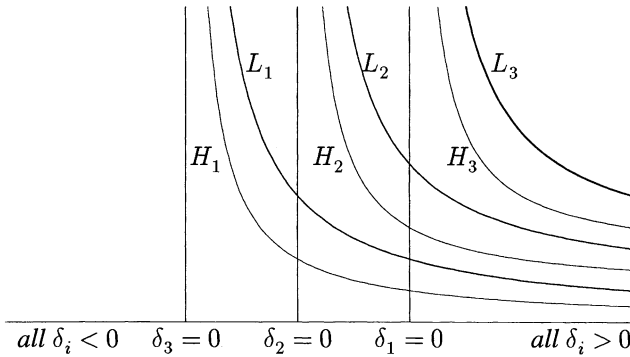
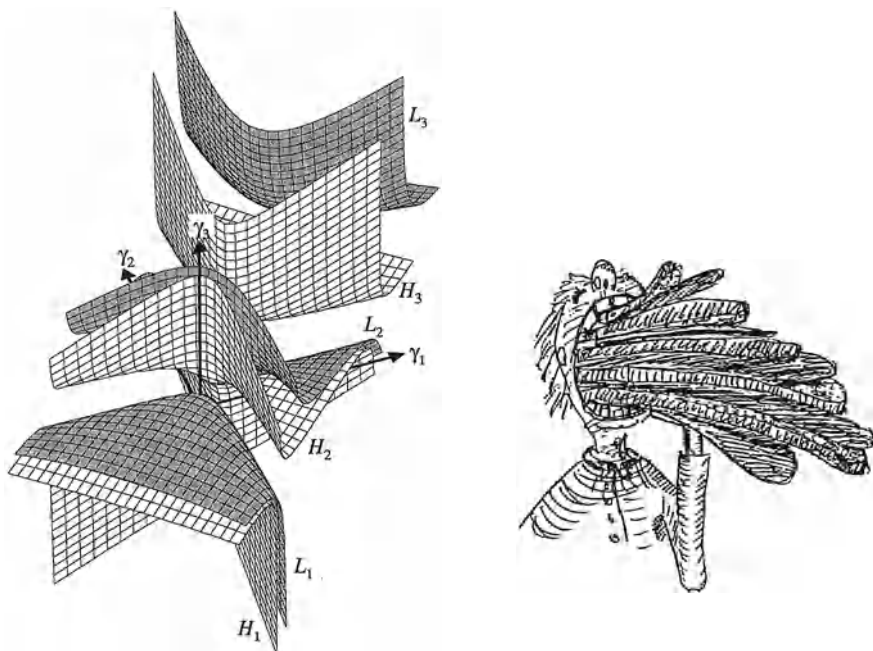


Fig. 4.7. Formation of the sandwich

Proof. When the δ_i have been chosen, inserting $\gamma_i = t\delta_i$ into (4.23) gives

$$\begin{aligned} \frac{1}{3!} - t \frac{\delta_1 + \delta_2 + \delta_3}{2!} + t^2 \frac{\delta_1 \delta_2 + \delta_1 \delta_3 + \delta_2 \delta_3}{1!} - t^3 \frac{\delta_1 \delta_2 \delta_3}{0!} &= 0 \\ \frac{1}{4!} - t \frac{\delta_1 + \delta_2 + \delta_3}{3!} + t^2 \frac{\delta_1 \delta_2 + \delta_1 \delta_3 + \delta_2 \delta_3}{2!} - t^3 \frac{\delta_1 \delta_2 \delta_3}{1!} &= 0 \end{aligned} \quad (4.26)$$

for L and H , respectively. These are third (in general s th) degree polynomials whose positive roots we have to study. We vary the δ 's, and hence the ray X , starting with all δ 's negative. The polynomials (4.26) then have all coefficients positive and obviously no positive real roots. When now *one* delta, say δ_3 , changes

Fig. 4.8. The sandwich for $s = 3 \dots$ and for $s = 5$

sign, the leading coefficients of (4.26) become zero and *one* root becomes infinite for each equation and satisfies asymptotically

$$\begin{aligned} \frac{\delta_1 \delta_2}{1!} - t \frac{\delta_1 \delta_2 \delta_3}{0!} &\approx 0 &\Rightarrow & t \approx \frac{1}{\delta_3} \\ \frac{\delta_1 \delta_2}{2!} - t \frac{\delta_1 \delta_2 \delta_3}{1!} &\approx 0 &\Rightarrow & t \approx \frac{1}{2\delta_3} \end{aligned} \quad (4.27)$$

for L and H , respectively. Thus H comes below and L comes above. Because of $L \cap H = \emptyset$ (4.22) these two roots can never cross and must therefore remain in this configuration (see Fig. 4.7).

When then successively δ_2 and δ_1 change sign, the same scene repeats itself again and again, always two sheets of H and L descend from above in that order and are layed on the lower sheets like slices of bread and ham of a giant sandwich. Because $L \cap H = \emptyset$, these sheets can never cross, two roots for L or H can never come together and become complex. So all roots must remain real and the theorem must be true.

A three-dimensional view of these surfaces is given in Fig. 4.8. □

The following theorem describes the form of the corresponding order star in all these sheets.

Theorem 4.24. *Let G_1, \dots, G_s be the open connected components of $\mathbb{R}^s \setminus H$ such that L_i lies in G_i , and let G_0 be the component containing the origin. Then the order star of $R(z)$ given by (4.20) possesses exactly k bounded fingers to the right of Cary Grant's part if and only if*

$$(\gamma_1, \dots, \gamma_s) \in G_k \cup H_k.$$

Proof. We prove this by a continuity argument letting the point $(\gamma_1, \dots, \gamma_s)$ travel through the sandwich. Since Cary Grant's part is always present (Remark 4.19), the number of bounded sectors can change only where the error constant C (4.21) changes sign, i.e., on the surfaces H_1, H_2, \dots, H_s . Fig. 4.9 gives some snap-shots from this voyage for $s = 3$ and $\gamma_1 = \gamma_2 = \gamma_3 = \gamma$. In this case the equations (4.23) become

$$\begin{aligned} \frac{1}{3!} - \frac{3\gamma}{2!} + \frac{3\gamma^2}{1!} - \frac{\gamma^3}{0!} &= 0 \\ \frac{1}{4!} - \frac{3\gamma}{3!} + \frac{3\gamma^2}{2!} - \frac{\gamma^3}{1!} &= 0 \end{aligned} \tag{4.28}$$

whose roots

$$\begin{aligned} \lambda_1 &= 0.158984, & \lambda_2 &= 0.435867, & \lambda_3 &= 2.40515 \\ \chi_1 &= 0.128886, & \chi_2 &= 0.302535, & \chi_3 &= 1.06858 \end{aligned} \tag{4.29}$$

do interlace nicely as required by Theorem 4.23. The affirmation of Theorem 4.24 for $s = 3$ can be clearly observed in Fig. 4.9.

For the proof of the general statement we also put $\gamma_1 = \dots = \gamma_s = \gamma$ and investigate the two extreme cases:

1. $\gamma = 0$: Here $R(z)$ is the Taylor polynomial $1 + z + \dots + z^s/s!$ whose order star has no bounded sector at all.

2. $\gamma \rightarrow \infty$: The numerator of $R(z)$ in (4.20) becomes for $s = 3$

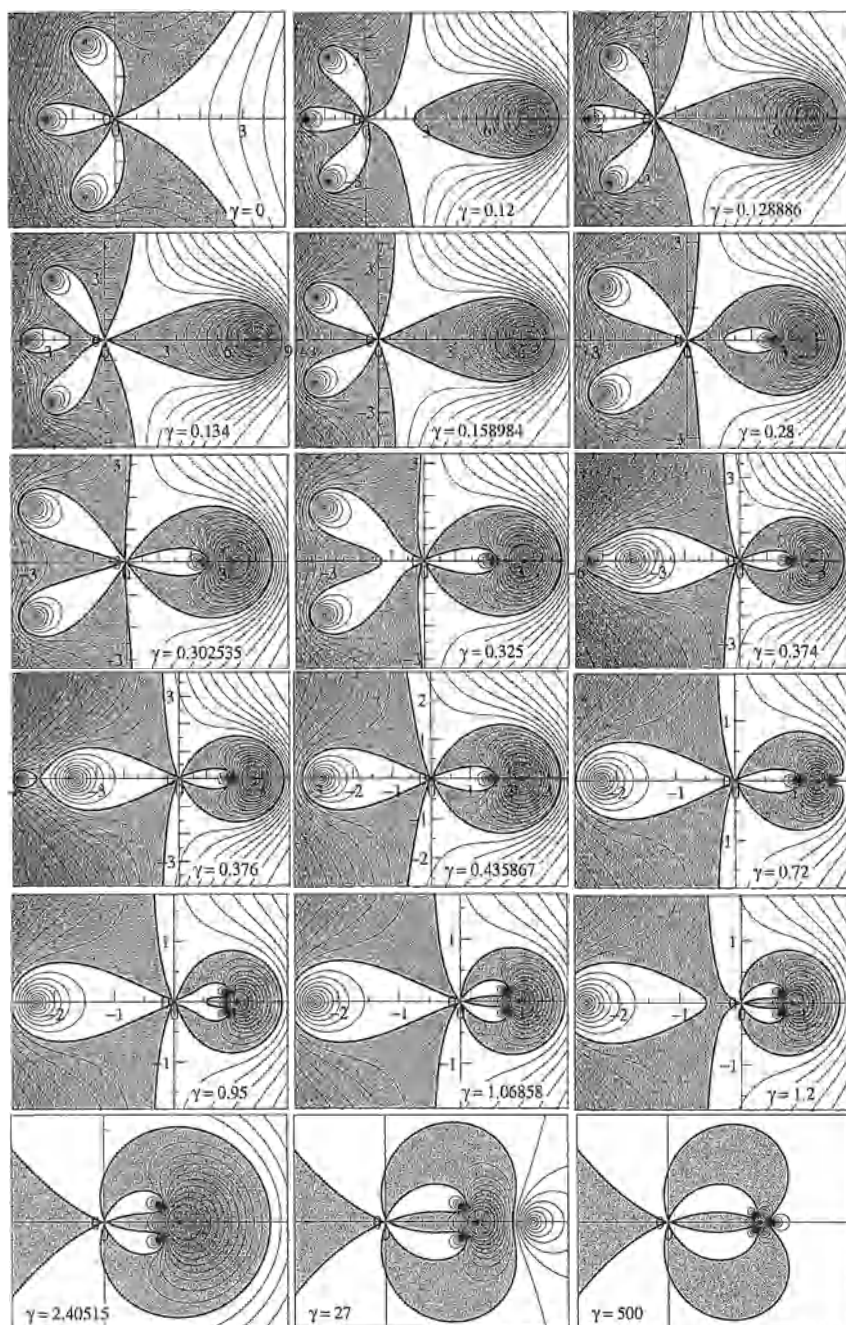
$$P(z) = 1 + z \left(\frac{1}{1!} - \frac{3\gamma}{0!} \right) + z^2 \left(\frac{1}{2!} - \frac{3\gamma}{1!} + \frac{3\gamma^2}{0!} \right) + z^3 \left(\frac{1}{3!} - \frac{3\gamma}{2!} + \frac{3\gamma^2}{1!} - \frac{\gamma^3}{0!} \right). \tag{4.30}$$

If we let $\gamma \rightarrow \infty$, this becomes with $z\gamma = w$

$$1 - w \left(3 + \mathcal{O}\left(\frac{1}{\gamma}\right) \right) + w^2 \left(3 + \mathcal{O}\left(\frac{1}{\gamma}\right) \right) - w^3 \left(1 + \mathcal{O}\left(\frac{1}{\gamma}\right) \right).$$

Therefore all roots $w_i \rightarrow 1$, hence $z_i \rightarrow 1/\gamma$ (see the last picture of Fig. 4.9). Therefore *no* zero of $R(z)$ can remain left of Cary Grant's part and we have s bounded fingers.

Since between these extreme cases, there are at most s crossings of the surface H , Theorem 4.24 must be true. \square

Fig. 4.9. Order stars for γ travelling through the sandwich

Theorem 4.25. *The function $R(z)$ defined by (4.20) can be I -stable only if*

$$(\gamma_1, \dots, \gamma_s) \in H_q \cup G_q \cup H_{q+1} \quad \text{if } s = 2q - 1$$

and

$$(\gamma_1, \dots, \gamma_s) \in G_q \cup H_{q+1} \cup G_{q+1} \quad \text{if } s = 2q.$$

Proof. The reason for this result is similar to Theorem 4.12. For I -stability the imaginary axis cannot intersect the order star and must therefore reach the origin through Cary Grant's part. Thus I -stability (and hence A -stability) is only possible (roughly) *in the middle* of the sandwich. Since at most $\lfloor (p+2)/2 \rfloor$ and at least $\lfloor (p+1)/2 \rfloor$ of the $p+1$ sectors of A start in \mathbb{C}^+ , the number k of bounded fingers satisfies

$$\left\lfloor \frac{p+2}{2} \right\rfloor \geq k \quad \text{and} \quad \left\lfloor \frac{p+1}{2} \right\rfloor \leq k.$$

Inserting $p = s + 1$ on H and $p = s$ on G we get the above results. \square

Multiple Real-Pole Approximations

... the next main result is obtained, saying that the least value of C is obtained when all the zeros of the denominator are equal (Nørsett & Wolfbrandt 1977)

Approximations for which all poles are equal, i.e., for which $\gamma_1 = \gamma_2 = \dots = \gamma_s = \gamma$ are called “multiple” real-pole approximations (Nørsett 1974). We again consider only approximations for which the order is $\geq s$. These satisfy, for $s = 3$,

$$R(z) = \frac{P(z)}{(1 - \gamma z)^3} \quad (4.31)$$

where $P(z)$ is given by (4.30), and their error constant is

$$C = \frac{1}{4!} - \frac{3\gamma}{3!} + \frac{3\gamma^2}{2!} - \frac{\gamma^3}{1!}. \quad (4.32)$$

Approximations with multiple poles have many computational advantages (the linear systems to be solved in Rosenbrock or DIRK methods have all the same matrix (see Sections IV.6 and IV.7)). We are now pleased to see that they also have the smallest error constants (Nørsett & Wolfbrandt 1977).

Theorem 4.26. *On each of the surfaces L_i and H_i ($i = 1, \dots, s$) the error constant C of (4.20) is minimized (in absolute value) when $\gamma_1 = \gamma_2 = \dots = \gamma_s$.*

Proof. Our proof uses relative order stars (similar to (4.13))

$$B = \left\{ z \in \mathbb{C} ; |q(z)| > 1 \right\}, \quad q(z) = \frac{R_{new}(z)}{R_{old}(z)}, \quad (4.33)$$

where $R_{old}(z)$ is a real-pole approximation of order $p = s + 1$ corresponding to $\gamma_1, \dots, \gamma_s$ and $R_{new}(z)$ is obtained by an infinitely small change of the γ 's. We assume that not all γ_i are identical and shall show that then the error constant can be decreased. After a permutation of the indices, we assume $\gamma_1 = \max(\gamma_i)$ (by Theorem 4.23 $\gamma_1 > 0$, so that $1/\gamma_1$ represents the pole on the positive real axis which is closest to the origin) and $\gamma_s < \gamma_1$. We don't allow arbitrary changes of the γ 's but we *decrease* γ_1 , keep $\gamma_2, \dots, \gamma_{s-1}$ fixed and determine γ_s by the defining equations for H (see (4.23)). For example, for $s = 3$ we have

$$\gamma_3 = \frac{\frac{1}{4!} - \frac{\gamma_1 + \gamma_2}{3!} + \frac{\gamma_1 \gamma_2}{2!}}{\frac{1}{3!} - \frac{\gamma_1 + \gamma_2}{2!} + \frac{\gamma_1 \gamma_2}{1!}}. \quad (4.34)$$

Since the poles and zeros of $R_{old}(z)$ depend continuously on the γ_i , poles and zeros of $q(z)$ appear always in pairs (we call them dipoles). By the maximum principle or by Remark 4.14, each boundary curve of B leaving the origin must lead to at least one dipole before it rejoins the origin. Since there are $s + 2 = p + 1$ dipoles of $q(z)$ (identical poles for $R_{old}(z)$ and $R_{new}(z)$ don't give rise to a dipole of $q(z)$) and $p + 1$ pairs of boundary curves of B leaving the origin (Remark 4.14), each such boundary curve passes through exactly one dipole before rejoining the origin. As a consequence no boundary curve of B can cross the real axis except at dipoles.

If the error constant of $R_{old}(z)$ satisfies $C_{old} < 0$, then, by Remark 4.20, $R_{old}(z)$ has no zero between $1/\gamma_1$ and the origin. Therefore also $q(z)$ possesses no dipole in this region. Since the pole of $R_{new}(z)$ is slightly larger than $1/\gamma_1$ (that of $R_{old}(z)$), the real axis between $1/\gamma_1$ and the origin must belong to the complement of B . Thus we have $C_{new} - C_{old} > 0$ by (4.14) and (4.15).

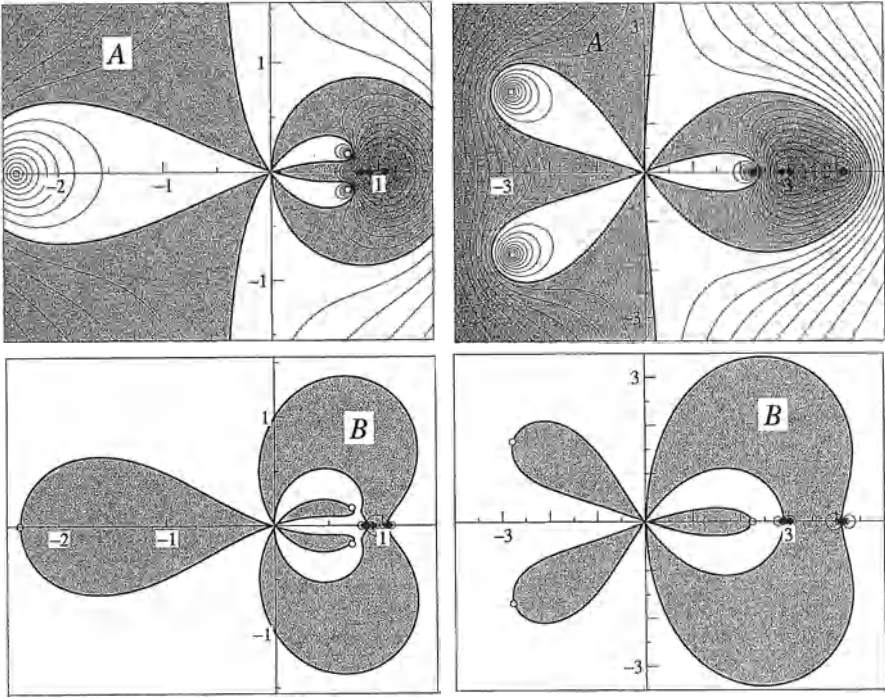
If $C_{old} > 0$ there is one additional dipole of $q(z)$ between $1/\gamma_1$ and the origin (see Remark 4.20). As above we conclude this time that $C_{new} - C_{old} < 0$.

In both cases $|C_{new}| < |C_{old}|$, since by continuity C_{new} is near to C_{old} . As a consequence no $(\gamma_1, \dots, \gamma_s) \in H$ with at least two different γ_i can minimize the error constant. As it becomes large in modulus when at least one γ_i tends to ∞ (this follows from Theorem 4.18 and from the fact that in this case $R(z)$ tends to an approximation with s replaced by $s - 1$) the minimal value of C must be attained when all poles are identical.

The proof for L is the same, there are only $s - 1$ zeros of $R(z)$ and the order is $p = s$. \square

An illustration of the order star B compared to A is given in Fig. 4.10. Another advantage of multiple real-pole approximations is exhibited by the following theorem:

Theorem 4.27 (Keeling 1989). *On each surface $H_i \cap \{(\gamma_1, \dots, \gamma_s); \gamma_j > 0\}$ the value $|R(\infty)|$ of (4.20) is minimized when $\gamma_1 = \gamma_2 = \dots = \gamma_s$.*



	left pictures: $C_{old} < 0$	right pictures: $C_{old} > 0$
R_{old}	$\gamma_1 = 1.2$ $\gamma_2 = 1.1$ $\gamma_3 = 0.9455446$	$\gamma_1 = 0.35$ $\gamma_2 = 0.33$ $\gamma_3 = 0.2406340$
R_{new}	$\gamma_1 = 1.17$ $\gamma_2 = 1.1$ $\gamma_3 = 0.9628661$	$\gamma_1 = 0.345$ $\gamma_2 = 0.33$ $\gamma_3 = 0.2440772$

Fig. 4.10. Order star A compared to B

Proof. The beginning of the proof is identical to that of Theorem 4.26. Besides $1/\gamma_1$ and $1/\gamma_s$ there is at best an even number of dipoles on the positive real axis to the right of $1/\gamma_1$. As in the proof above we conclude that a right-neighbourhood of $1/\gamma_1$ belongs to B so that ∞ must lie in its complement (cf. Fig. 4.10). This implies

$$|R_{new}(\infty)| < |R_{old}(\infty)|$$

As a consequence no element of $H \cap \{(\gamma_1, \dots, \gamma_s); \gamma_j > 0\}$ with at least two γ_j different can minimize $|R(\infty)|$. Also $|R(\infty)|$ increases if $\gamma_1 \rightarrow \infty$. The statement now follows from the fact that $|R(\infty)|$ tends to infinity when at least one γ_j approaches zero. \square

Exercises

1. (Ehle 1968). Compute the polynomial $E(y)$ for the third and fourth Padé subdiagonal $R_{k,k+3}(z)$ and $R_{k,k+4}(z)$ (which, by Proposition 3.4 consists of two terms only). Show that these approximations violate (3.6) and cannot be A -stable.

2. Prove the general formula

$$E(y) = \left(\frac{k!}{(k+j)!} \right)^2 \sum_{r=\lceil (k+j+2)/2 \rceil}^j \frac{(-1)^{j-r}}{(j-r)!} \left(\prod_{q=1}^{j-r} (j-q+1)(k+q)(r-k-q) \right) y^{2r}$$

for the Padé approximations R_{kj} ($j \geq k$).

3. (For the fans of mathematical precision). Derive the following formulas for the roots λ_i and χ_i of (4.28)

$$\begin{aligned} \chi_1 &= \frac{1}{2} + \frac{1}{\sqrt{3}} \cos \frac{13\pi}{18}, & \lambda_1 &= 1 + \sqrt{2} \cos \left(\frac{\theta + 2\pi}{3} \right), \\ \chi_2 &= \frac{1}{2} + \frac{1}{\sqrt{3}} \cos \frac{25\pi}{18}, & \lambda_2 &= 1 + \sqrt{2} \cos \left(\frac{\theta + 4\pi}{3} \right), \\ \chi_3 &= \frac{1}{2} + \frac{1}{\sqrt{3}} \cos \frac{\pi}{18}, & \lambda_3 &= 1 + \sqrt{2} \cos \left(\frac{\theta}{3} \right), \end{aligned}$$

where $\theta = \arctan(\sqrt{2}/4)$.

Hint. Use the Cardano-Viète formula (e.g., Hairer & Wanner (1995), page 66).

4. Prove that all zeros of

$$\frac{x^s}{s!} - S_1 \frac{x^{s-1}}{(s-1)!} + S_2 \frac{x^{s-2}}{(s-2)!} - \dots \pm S_s$$

are real and distinct whenever all zeros of

$$Q(z) = 1 - zS_1 + z^2S_2 - \dots \pm z^sS_s, \quad S_s \neq 0$$

are real. Also, both polynomials have the same number of positive (and negative) zeros (Nørsett & Wanner 1979, Bales, Karakashian & Serbin 1988).

Hint. Apply Theorem 4.23. This furnishes a geometric proof of a classical result (see e.g., Pólya & Szegő (1925), Volume II, Part V, No.65) and allows us to interpret $R(z)$ as the stability function of a (real) collocation method.

5. Prove that $(\gamma, \dots, \gamma) \in L$ (Definition 4.21) if and only if $L_s(1/\gamma) = 0$, where $L_s(x)$ denotes the *Laguerre polynomial of degree s* (see Abramowitz & Stegun (1964), Formula 22.3.9 or Formula (6.11) below).

IV.5 Construction of Implicit Runge-Kutta Methods

Although most of these methods appear at the moment to be largely of theoretical interest . . . (B.L. Ehle 1968)

In Sect. II.7 the first implicit Runge-Kutta methods were introduced. As we saw in Sect. IV.3, not all of them are suitable for the solution of stiff differential equations. This section is devoted to the collection of several classes of fully implicit Runge-Kutta methods possessing good stability properties.

The construction of such methods relies heavily on the simplifying assumptions

$$\begin{aligned} B(p) : \quad & \sum_{i=1}^s b_i c_i^{q-1} = \frac{1}{q} \quad q = 1, \dots, p; \\ C(\eta) : \quad & \sum_{j=1}^s a_{ij} c_j^{q-1} = \frac{c_i^q}{q} \quad i = 1, \dots, s, \quad q = 1, \dots, \eta; \\ D(\zeta) : \quad & \sum_{i=1}^s b_i c_i^{q-1} a_{ij} = \frac{b_j}{q} (1 - c_j^q) \quad j = 1, \dots, s, \quad q = 1, \dots, \zeta. \end{aligned}$$

Condition $B(p)$ simply means that the quadrature formula (b_i, c_i) is of order p . The importance of the other two conditions is seen from the following fundamental theorem, which was derived in Sect. II.7.

Theorem 5.1 (Butcher 1964). *If the coefficients b_i, c_i, a_{ij} of a Runge-Kutta method satisfy $B(p), C(\eta), D(\zeta)$ with $p \leq \eta + \zeta + 1$ and $p \leq 2\eta + 2$, then the method is of order p .* \square

Gauss Methods

These processes, named “Kuntzmann-Butcher methods” in Sect. II.7, are collocation methods based on the Gaussian quadrature formulas, i.e., c_1, \dots, c_s are the zeros of the shifted Legendre polynomial of degree s ,

$$\frac{d^s}{dx^s} (x^s (x-1)^s).$$

For the sake of completeness we present the first of these in Tables 5.1 and 5.2.

Table 5.1. Gauss methods of order 2 and 4

$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2} - \frac{\sqrt{3}}{6}$	$\frac{1}{4}$	$\frac{1}{4} - \frac{\sqrt{3}}{6}$
		$\frac{1}{2} + \frac{\sqrt{3}}{6}$	$\frac{1}{4} + \frac{\sqrt{3}}{6}$	$\frac{1}{4}$
	1		$\frac{1}{2}$	$\frac{1}{2}$

Table 5.2. Gauss method of order 6

$\frac{1}{2} - \frac{\sqrt{15}}{10}$	$\frac{5}{36}$	$\frac{2}{9} - \frac{\sqrt{15}}{15}$	$\frac{5}{36} - \frac{\sqrt{15}}{30}$
$\frac{1}{2}$	$\frac{5}{36} + \frac{\sqrt{15}}{24}$	$\frac{2}{9}$	$\frac{5}{36} - \frac{\sqrt{15}}{24}$
$\frac{1}{2} + \frac{\sqrt{15}}{10}$	$\frac{5}{36} + \frac{\sqrt{15}}{30}$	$\frac{2}{9} + \frac{\sqrt{15}}{15}$	$\frac{5}{36}$
	$\frac{5}{18}$	$\frac{4}{9}$	$\frac{5}{18}$

Theorem 5.2 (Butcher 1964, Ehle 1968). *The s -stage Gauss method is of order $2s$. Its stability function is the (s, s) -Padé approximation and the method is A-stable.*

Proof. The order result has already been proved in Sect. II.7. Since the degrees of the numerator and the denominator are not larger than s for any s -stage Runge-Kutta method, the stability function of this $2s$ -order method is the (s, s) -Padé approximation by Theorem 3.11. A-stability thus follows from Theorem 4.12. \square

Radau IA and Radau IIA Methods

Butcher (1964) introduced Runge-Kutta methods based on the Radau and Lobatto quadrature formulas. He called them processes of type I, II or III according to whether c_1, \dots, c_s are the zeros of

$$\text{I: } \frac{d^{s-1}}{dx^{s-1}} \left(x^s (x-1)^{s-1} \right), \quad (\text{Radau left}) \quad (5.1)$$

$$\text{II: } \frac{d^{s-1}}{dx^{s-1}} \left(x^{s-1} (x-1)^s \right), \quad (\text{Radau right}) \quad (5.2)$$

$$\text{III: } \frac{d^{s-2}}{dx^{s-2}} \left(x^{s-1} (x-1)^{s-1} \right). \quad (\text{Lobatto}) \quad (5.3)$$

The weights b_1, \dots, b_s are chosen such that the quadrature formula satisfies $B(s)$, which implies $B(2s-1)$ in the Radau case and $B(2s-2)$ in the Lobatto case

(see Lemma 5.15 below). Unfortunately, none of these methods of Butcher turned out to be A -stable (see e.g. Table 3.1). Ehle (1969) took up the ideas of Butcher and constructed methods of type I, II and III with excellent stability properties. Independently, Axelsson (1969) found the Radau IIA methods together with an elegant proof of their A -stability.

The s -stage Radau IA method is of type I, where the coefficients a_{ij} , ($i, j = 1, \dots, s$) are defined by condition $D(s)$. This is uniquely possible since the c_i are distinct and the b_i not zero. Tables 5.3 and 5.4 present the first of these methods.

Table 5.3. Radau IA methods of orders 1 and 3

0	1
	1

0	$\frac{1}{4}$	$-\frac{1}{4}$
$\frac{2}{3}$	$\frac{1}{4}$	$\frac{5}{12}$
	$\frac{1}{4}$	$\frac{3}{4}$

Table 5.4. Radau IA method of order 5

0	$\frac{1}{9}$	$\frac{-1 - \sqrt{6}}{18}$	$\frac{-1 + \sqrt{6}}{18}$
$\frac{6 - \sqrt{6}}{10}$	$\frac{1}{9}$	$\frac{88 + 7\sqrt{6}}{360}$	$\frac{88 - 43\sqrt{6}}{360}$
$\frac{6 + \sqrt{6}}{10}$	$\frac{1}{9}$	$\frac{88 + 43\sqrt{6}}{360}$	$\frac{88 - 7\sqrt{6}}{360}$
	$\frac{1}{9}$	$\frac{16 + \sqrt{6}}{36}$	$\frac{16 - \sqrt{6}}{36}$

Ehle's type II processes are obtained by imposing condition $C(s)$. By Theorem II.7.7 this results in the collocation methods based on the zeros of (5.2). They are called Radau IIA methods. Examples are given in Tables 5.5 and 5.6. For $s = 1$ we obtain the implicit Euler method.

Theorem 5.3. *The s -stage Radau IA method and the s -stage Radau IIA method are of order $2s - 1$. Their stability function is the $(s - 1, s)$ subdiagonal Padé approximation. Both methods are A -stable.*

Proof. The stated orders follow from Theorem 5.1 and Lemma 5.4 below. Since $c_1 = 0$ for the Radau IA method, $D(s)$ with $j = 1$ and $B(2s - 1)$ imply (3.14). Similarly, for the Radau IIA method, $c_s = 1$ and $C(s)$ imply (3.13). Therefore, in both cases, the numerator of the stability function is of degree $\leq s - 1$ by Proposition 3.8. The statement now follows from Theorem 3.11 and Theorem 4.12.

□

Table 5.5. Radau IIA methods of orders 1 and 3

		$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	1	1	$\frac{3}{4}$	$\frac{1}{4}$
	1		$\frac{3}{4}$	$\frac{1}{4}$

Table 5.6. Radau IIA method of order 5

$\frac{4 - \sqrt{6}}{10}$	$\frac{88 - 7\sqrt{6}}{360}$	$\frac{296 - 169\sqrt{6}}{1800}$	$\frac{-2 + 3\sqrt{6}}{225}$
$\frac{4 + \sqrt{6}}{10}$	$\frac{296 + 169\sqrt{6}}{1800}$	$\frac{88 + 7\sqrt{6}}{360}$	$\frac{-2 - 3\sqrt{6}}{225}$
1	$\frac{16 - \sqrt{6}}{36}$	$\frac{16 + \sqrt{6}}{36}$	$\frac{1}{9}$
	$\frac{16 - \sqrt{6}}{36}$	$\frac{16 + \sqrt{6}}{36}$	$\frac{1}{9}$

Lemma 5.4. Let an s -stage Runge-Kutta method have distinct c_1, \dots, c_s and non-zero weights b_1, \dots, b_s . Then we have

- a) $C(s)$ and $B(s + \nu)$ imply $D(\nu)$;
b) $D(s)$ and $B(s + \nu)$ imply $C(\nu)$.

Proof. Put

$$d_j^{(q)} := \sum_{i=1}^s b_i c_i^{q-1} a_{ij} - \frac{b_j}{q} (1 - c_j^q). \quad (5.4)$$

Conditions $C(s)$ and $B(s + \nu)$ imply

$$\sum_{j=1}^s d_j^{(q)} c_j^{k-1} = 0 \quad \text{for } k = 1, \dots, s \text{ and } q = 1, \dots, \nu.$$

The vector $(d_1^{(q)}, \dots, d_s^{(q)})$ must vanish, because it is the solution of a homogeneous linear system with a non singular matrix of Vandermonde type. This proves $D(\nu)$.

For part (b) one defines

$$e_i^{(q)} := \sum_{j=1}^s a_{ij} c_j^{q-1} - \frac{c_i^q}{q}$$

and applies a similar argument to

$$\sum_{i=1}^s b_i c_i^{k-1} e_i^{(q)} = 0, \quad k = 1, \dots, s, \quad q = 1, \dots, \nu. \quad \square$$

Lobatto IIIA, IIIB and IIIC Methods

For all type III processes the c_i are the zeros of the polynomial (5.3) and the weights b_i are such that $B(2s-2)$ is satisfied.

The coefficients a_{ij} are defined by $C(s)$ for the Lobatto IIIA methods. It is therefore a collocation method. For the Lobatto IIIB methods we impose $D(s)$ and, finally, for the Lobatto IIIC methods we put

$$a_{i1} = b_1 \quad \text{for } i = 1, \dots, s \quad (5.5)$$

and determine the remaining a_{ij} by $C(s-1)$. Ehle (1969) introduced the first two classes, and presented the IIIC methods for $s \leq 3$. The general definition of the IIIC methods is due to Chipman (1971); see also Axelsson (1972). Examples are given in Tables 5.7-5.12.

Table 5.7. Lobatto IIIA methods of orders 2 and 4

			0	0	0
0	0	0	$\frac{1}{2}$	$\frac{5}{24}$	$-\frac{1}{24}$
1	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{6}$	$\frac{1}{6}$
	$\frac{1}{2}$	$\frac{1}{2}$		$\frac{1}{6}$	$\frac{1}{6}$

Table 5.8. Lobatto IIIA method of order 6

0	0	0	0	0
$\frac{5-\sqrt{5}}{10}$	$\frac{11+\sqrt{5}}{120}$	$\frac{25-\sqrt{5}}{120}$	$\frac{25-13\sqrt{5}}{120}$	$\frac{-1+\sqrt{5}}{120}$
$\frac{5+\sqrt{5}}{10}$	$\frac{11-\sqrt{5}}{120}$	$\frac{25+13\sqrt{5}}{120}$	$\frac{25+\sqrt{5}}{120}$	$\frac{-1-\sqrt{5}}{120}$
1	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$
	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$

Theorem 5.5. *The s -stage Lobatto IIIA, IIIB and IIIC methods are of order $2s-2$. The stability function for the Lobatto IIIA and IIIB methods is the diagonal $(s-1, s-1)$ -Padé approximation. For the Lobatto IIIC method it is the $(s-2, s)$ -Padé approximation. All these methods are A -stable.*

Proof. We first prove that the IIIC methods satisfy $D(s-1)$. Condition (5.5) implies $d_1^{(q)} = 0$ ($q = 1, \dots, s-1$) for $d_1^{(q)}$ given by (5.4). Conditions $C(s-1)$

Table 5.9. Lobatto IIIB methods of orders 2 and 4

			0	$\frac{1}{6}$	$-\frac{1}{6}$	0
0	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}$	0
1	$\frac{1}{2}$	0	1	$\frac{1}{6}$	$\frac{5}{6}$	0
	$\frac{1}{2}$	$\frac{1}{2}$		$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Table 5.10. Lobatto IIIB method of order 6

0	$\frac{1}{12}$	$\frac{-1-\sqrt{5}}{24}$	$\frac{-1+\sqrt{5}}{24}$	0
$\frac{5-\sqrt{5}}{10}$	$\frac{1}{12}$	$\frac{25+\sqrt{5}}{120}$	$\frac{25-13\sqrt{5}}{120}$	0
$\frac{5+\sqrt{5}}{10}$	$\frac{1}{12}$	$\frac{25+13\sqrt{5}}{120}$	$\frac{25-\sqrt{5}}{120}$	0
1	$\frac{1}{12}$	$\frac{11-\sqrt{5}}{24}$	$\frac{11+\sqrt{5}}{24}$	0
	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$

Table 5.11. Lobatto IIIC methods of orders 2 and 4

			0	$\frac{1}{6}$	$-\frac{1}{3}$	$\frac{1}{6}$
0	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{1}{2}$	$\frac{1}{2}$		$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

Table 5.12. Lobatto IIIC method of order 6

0	$\frac{1}{12}$	$\frac{-\sqrt{5}}{12}$	$\frac{\sqrt{5}}{12}$	$\frac{-1}{12}$
$\frac{5-\sqrt{5}}{10}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{10-7\sqrt{5}}{60}$	$\frac{\sqrt{5}}{60}$
$\frac{5+\sqrt{5}}{10}$	$\frac{1}{12}$	$\frac{10+7\sqrt{5}}{60}$	$\frac{1}{4}$	$\frac{-\sqrt{5}}{60}$
1	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$
	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$

and $B(2s - 2)$ then yield

$$\sum_{j=2}^s d_j^{(q)} c_j^{k-1} = 0 \quad \text{for } k = 1, \dots, s-1 \text{ and } q = 1, \dots, s-1.$$

As in the proof of Lemma 5.4 we deduce $D(s-1)$. All order statements now follow from Lemma 5.4 and Theorem 5.1.

By definition, the first row of the Runge-Kutta matrix A vanishes for the IIIA methods, and its last column vanishes for the IIIB methods. The denominator of the stability function is therefore of degree $\leq s-1$. Similarly, the last row of $A - \mathbb{1}b^T$ vanishes for IIIA, and the first column of $A - \mathbb{1}b^T$ for IIIB. Therefore, the numerator of the stability function is also of degree $\leq s-1$ by Formula (3.3). It now follows from Theorem 3.11 that both methods have the $(s-1, s-1)$ -Padé approximation as stability function.

For the IIIC process the first column as well as the last row of $A - \mathbb{1}b^T$ vanish. Thus the degree of the numerator of the stability function is at most $s-2$ by Formula (3.3). Again, Theorem 3.11 and Theorem 4.12 imply the statement. \square

For a summary of these statements see Table 5.13.

Table 5.13. Fully implicit Runge-Kutta methods

method	simplifying assumptions			order	stability function
Gauss	$B(2s)$	$C(s)$	$D(s)$	$2s$	(s, s) -Padé
Radau IA	$B(2s-1)$	$C(s-1)$	$D(s)$	$2s-1$	$(s-1, s)$ -Padé
Radau IIA	$B(2s-1)$	$C(s)$	$D(s-1)$	$2s-1$	$(s-1, s)$ -Padé
Lobatto IIIA	$B(2s-2)$	$C(s)$	$D(s-2)$	$2s-2$	$(s-1, s-1)$ -Padé
Lobatto IIIB	$B(2s-2)$	$C(s-2)$	$D(s)$	$2s-2$	$(s-1, s-1)$ -Padé
Lobatto IIIC	$B(2s-2)$	$C(s-1)$	$D(s-1)$	$2s-2$	$(s-2, s)$ -Padé

The W -Transformation

We now attack the explicit construction of all Runge-Kutta methods covered by Theorem 5.1. The first observation is (Chipman 1971, Burrage 1978) that $C(\eta)$ can be written as

$$\begin{pmatrix} a_{11} & \dots & a_{1s} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ a_{s1} & \dots & a_{ss} \end{pmatrix} \begin{pmatrix} 1 & c_1 & \dots & c_1^{\eta-1} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & c_s & \dots & c_s^{\eta-1} \end{pmatrix} = \begin{pmatrix} 1 & c_1 & \dots & c_1^\eta \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & c_s & \dots & c_s^\eta \end{pmatrix} \begin{pmatrix} 0 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ 0 & \frac{1}{2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\eta} \end{pmatrix}. \quad (5.6)$$

Hence, if V is the Vandermonde matrix

$$V = \begin{pmatrix} 1 & c_1 & \dots & c_1^{s-1} \\ \vdots & \vdots & & \vdots \\ 1 & c_s & \dots & c_s^{s-1} \end{pmatrix},$$

then the first η (for $\eta \leq s-1$) columns of $V^{-1}AV$ must have the special structure (with many zeros) of the rightmost matrix in (5.6). This “ V -transformation” already considerably simplifies the discussion of order and stability of methods governed by $C(\eta)$ with η close to s (Burrage 1978). Thus, *collocation methods* ($\eta = s$) are characterized by

$$V^{-1}AV = \begin{pmatrix} 0 & & & & -\varrho_0/s \\ 1 & 0 & & & -\varrho_1/s \\ & 1/2 & 0 & & -\varrho_2/s \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & 0 \\ & & & & 1/(s-1) & -\varrho_{s-1}/s \end{pmatrix} \quad (5.7)$$

where the ϱ 's are the coefficients of $M(t) = \prod_{i=1}^s (t - c_i)$ and appear when the c_i^s in (5.6) are replaced by lower powers. Whenever some of the columns of $V^{-1}AV$ are *not* as in (5.7), a nice idea of Nørsett allows one to interpret the method as a *perturbed collocation* method (see Nørsett & Wanner (1981) for more details).

However, the V -transformation has some drawbacks: it does not allow a similar characterization of $D(\zeta)$, and the discussions of A - and B -stability remain fairly complicated (see e.g. the above cited papers). It was then discovered (Hairer & Wanner 1981, 1982) that nicer results are obtained, if the Vandermonde matrix V is replaced by a matrix W whose elements are *orthogonal polynomials* evaluated at c_i . We therefore use the (non standard) notation

$$P_k(x) = \frac{\sqrt{2k+1}}{k!} \frac{d^k}{dx^k} (x^k(x-1)^k) = \sqrt{2k+1} \sum_{j=0}^k (-1)^{j+k} \binom{k}{j} \binom{j+k}{j} x^j \quad (5.8)$$

for the *shifted Legendre polynomials* normalized so that

$$\int_0^1 P_k^2(x) dx = 1. \quad (5.9)$$

These polynomials satisfy the integration formulas

$$\begin{aligned} \int_0^x P_0(t) dt &= \xi_1 P_1(x) + \frac{1}{2} P_0(x) \\ \int_0^x P_k(t) dt &= \xi_{k+1} P_{k+1}(x) - \xi_k P_{k-1}(x) \quad k = 1, 2, \dots \end{aligned} \quad (5.10)$$

with

$$\xi_k = \frac{1}{2\sqrt{4k^2 - 1}} \quad (5.11)$$

(Exercise 1). Instead of (5.7) we now have the following result.

Theorem 5.6. *Let W be defined by*

$$w_{ij} = P_{j-1}(c_i), \quad i = 1, \dots, s, \quad j = 1, \dots, s, \quad (5.12)$$

and let A be the coefficient matrix for the Gauss method of order $2s$. Then,

$$W^{-1}AW = \begin{pmatrix} 1/2 & -\xi_1 & & & \\ \xi_1 & 0 & -\xi_2 & & \\ & \xi_2 & \ddots & \ddots & \\ & & \ddots & 0 & -\xi_{s-1} \\ & & & \xi_{s-1} & 0 \end{pmatrix} =: X_G. \quad (5.13)$$

Proof. We first write $C(\eta)$ in the form

$$\sum_{j=1}^s a_{ij} p(c_j) = \int_0^{c_i} p(x) dx \quad \text{if } \deg(p) \leq \eta - 1, \quad (5.14)$$

which, by (5.10), is equivalent to

$$\begin{aligned} \sum_{j=1}^s a_{ij} P_0(c_j) &= \xi_1 P_1(c_i) + \frac{1}{2} P_0(c_i) \\ \sum_{j=1}^s a_{ij} P_k(c_j) &= \xi_{k+1} P_{k+1}(c_i) - \xi_k P_{k-1}(c_i) \quad k = 1, \dots, \eta - 1. \end{aligned} \quad (5.15)$$

For $\eta = s$, inserting (5.12), and using matrix notation, this becomes

$$\begin{pmatrix} a_{11} & \dots & a_{1s} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ a_{s1} & \dots & a_{ss} \end{pmatrix} \begin{pmatrix} w_{11} & \dots & w_{1s} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ w_{s1} & \dots & w_{ss} \end{pmatrix} = \begin{pmatrix} w_{11} & \dots & w_{1s} & P_s(c_1) \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & \vdots \\ w_{s1} & \dots & w_{ss} & P_s(c_s) \end{pmatrix} \begin{pmatrix} 1/2 & -\xi_1 & & & \\ \xi_1 & 0 & -\xi_2 & & \\ & \xi_2 & \ddots & \ddots & \\ & & \ddots & 0 & -\xi_{s-1} \\ & & & \xi_{s-1} & 0 \end{pmatrix}. \quad (5.16)$$

Since for the Gauss processes we have $P_s(c_1) = \dots = P_s(c_s) = 0$, the last column respectively row of the right hand matrices can be dropped and we obtain (5.13). \square

In what follows we shall study similar results for other implicit Runge-Kutta methods. We first formulate the following lemma, which is an immediate consequence of (5.15) and (5.16).

Lemma 5.7. *Let A be the coefficient matrix of an implicit Runge-Kutta method and let W be a nonsingular matrix with*

$$w_{ij} = P_{j-1}(c_i) \quad \text{for } i = 1, \dots, s, \quad j = 1, \dots, \eta + 1.$$

Then $C(\eta)$ (with $\eta \leq s - 1$) is equivalent to the fact that the first η columns of $W^{-1}AW$ are equal to those of X_G in (5.13). \square

The second type of simplifying assumption, $D(\zeta)$, is now written in the form

$$\sum_{i=1}^s b_i p(c_i) a_{ij} = b_j \int_{c_j}^1 p(x) dx \quad \text{if } \deg(p) \leq \zeta - 1. \quad (5.17)$$

The integration formulas (5.10) together with orthogonality relations

$$\int_0^1 P_0(x) dx = 1, \quad \int_0^1 P_k(x) dx = \int_0^1 P_0(x) P_k(x) dx = 0 \quad \text{for } k = 1, 2, \dots$$

show that $D(\zeta)$ (i.e., (5.17)) is equivalent to

$$\begin{aligned} \sum_{i=1}^s P_0(c_i) b_i a_{ij} &= \left(\frac{1}{2} P_0(c_j) - \xi_1 P_1(c_j) \right) b_j \\ \sum_{i=1}^s P_k(c_i) b_i a_{ij} &= \left(\xi_k P_{k-1}(c_j) - \xi_{k+1} P_{k+1}(c_j) \right) b_j \quad k = 1, \dots, \zeta - 1. \end{aligned} \quad (5.18)$$

This can be stated as

Lemma 5.8. *As in the preceding lemma, let W be a nonsingular matrix with*

$$w_{ij} = P_{j-1}(c_i) \quad \text{for } i = 1, \dots, s, \quad j = 1, \dots, \zeta + 1,$$

and let $B = \text{diag}(b_1, \dots, b_s)$ with $b_i \neq 0$. Then $D(\zeta)$ (with $\zeta \leq s - 1$) is equivalent to the condition that the first ζ rows of the matrix $(W^T B)A(W^T B)^{-1}$ are equal to those of X_G in (5.13) (if B is singular, we still have (5.19) below).

Proof. Formulas (5.18), written in matrix form, give

$$W^T B A = \begin{pmatrix} 1/2 & -\xi_1 & & & \\ \xi_1 & 0 & \ddots & & \\ & \ddots & \ddots & -\xi_{\zeta-1} & \\ & & \xi_{\zeta-1} & 0 & -\xi_{\zeta} \\ * & * & \ddots & \ddots & \ddots & * \\ * & * & \ddots & \ddots & \ddots & * \end{pmatrix} W^T B. \quad (5.19)$$

\square

It is now a natural and interesting question, whether both transformation matrices of the foregoing lemmas can be made equal, i.e., whether

$$W^T B = W^{-1} \quad \text{or} \quad W^T B W = I. \quad (5.20)$$

A first result is:

Lemma 5.9. *For any quadrature formula of order $\geq 2s - 1$ the matrix*

$$W = \left(P_{j-1}(c_i) \right)_{i,j=1,\dots,s} \quad (5.21)$$

satisfies (5.20).

Proof. If the quadrature formula is of sufficiently high order, the polynomials $P_k(x)P_l(x)$ ($k + l \leq 2s - 2$) are integrated exactly, i.e.,

$$\sum_{i=1}^s b_i P_k(c_i) P_l(c_i) = \int_0^1 P_k(x) P_l(x) dx = \delta_{kl}; \quad (5.22)$$

this, however, is simply $W^T B W = I$. \square

Unfortunately, Condition (5.20) is too restrictive for many methods. We therefore relax our requirements as follows:

Definition 5.10. Let η, ζ be given integers between 0 and $s - 1$. We say that an $s \times s$ -matrix W satisfies $T(\eta, \zeta)$ for the quadrature formula $(b_i, c_i)_{i=1}^s$ if

$$\left. \begin{array}{l} \text{a) } W \text{ is nonsingular} \\ \text{b) } w_{ij} = P_{j-1}(c_i) \quad i = 1, \dots, s, \quad j = 1, \dots, \max(\eta, \zeta) + 1 \\ \text{c) } W^T B W = \begin{pmatrix} I & 0 \\ 0 & R \end{pmatrix} \end{array} \right\} T(\eta, \zeta)$$

where I is the $(\zeta + 1) \times (\zeta + 1)$ identity matrix; R is an arbitrary $(s - \zeta - 1) \times (s - \zeta - 1)$ matrix.

The main result is given in the following theorem. Together with Theorem 5.1 it is very helpful for the construction of high order methods (see Examples 5.16 and 5.24, and Theorem 13.15).

Theorem 5.11. *Let W satisfy $T(\eta, \zeta)$ for the quadrature formula $(b_i, c_i)_{i=1}^s$. Then for a Runge-Kutta method based on (b_i, c_i) we have, for the matrix $X = W^{-1} A W$,*

$$\left. \begin{array}{l} \text{a) the first } \eta \text{ columns of } X \text{ are those of } X_G \iff C(\eta), \\ \text{b) the first } \zeta \text{ rows of } X \text{ are those of } X_G \iff D(\zeta). \end{array} \right\}$$

Proof. The equivalence of (a) with $C(\eta)$ follows from Lemma 5.7. For the proof of (b) we multiply (5.19) from the right by W and obtain

$$W^T B W \cdot X = \tilde{X} \cdot W^T B W$$

where \tilde{X} is the large matrix of (5.19). Because of Condition (c) of $T(\eta, \zeta)$ the first ζ rows of \tilde{X} and X must be the same (write them as block matrices). The statement now follows from Lemma 5.8. \square

We have still left open the question of the existence of W satisfying $T(\eta, \zeta)$. The following two lemmas and Theorem 5.14 give an answer.

Lemma 5.12. *If the quadrature formula has distinct nodes c_i and all weights positive ($b_i > 0$) and if it is of order p with $p \geq 2\eta + 1$ and $p \geq 2\zeta + 1$, then the matrix*

$$W = \left(p_{j-1}(c_i) \right)_{i,j=1,\dots,s} \quad (5.23)$$

possesses property $T(\eta, \zeta)$ and satisfies (5.20). Here $p_j(x)$ is the polynomial of degree j orthonormalized for the scalar product

$$\langle p, r \rangle = \sum_{i=1}^s b_i p(c_i) r(c_i). \quad (5.24)$$

Proof. The positivity of the b 's makes (5.24) a scalar product on the space of polynomials of degree $\leq s-1$. Because of the order property (compare with (5.22)), the orthonormalized $p_j(x)$ must coincide for $j \leq \max(\eta, \zeta)$ with the Legendre polynomials $P_j(x)$. Orthonormality with respect to (5.24) means that $W^T B W = I$. \square

Lemma 5.13. *If the quadrature formula has distinct nodes c_i and is of order $p \geq s + \zeta$, then W defined by (5.21) has property $T(\eta, \zeta)$.*

Proof. Because of $p \geq s + \zeta$, (5.22) holds for $k = 0, \dots, s-1$ and $l = 0, \dots, \zeta$. This ensures (c) of Definition 5.10. \square

Theorem 5.14. *Let the quadrature formula be of order p . Then there exists a transformation with property $T(\eta, \zeta)$ if and only if*

$$p \geq \eta + \zeta + 1 \quad \text{and} \quad p \geq 2\zeta + 1, \quad (5.25)$$

and at least $\max(\eta, \zeta) + 1$ numbers among c_1, \dots, c_s are distinct.

Proof. Set $\nu = \max(\eta, \zeta)$ and denote the columns of the transformation W by w_1, \dots, w_s . In virtue of (b) of $T(\eta, \zeta)$ we have

$$w_j = \left(P_{j-1}(c_1), \dots, P_{j-1}(c_s) \right)^T \quad \text{for } j = 1, \dots, \nu + 1.$$

These $\nu + 1$ columns are linearly independent only if at least $\nu + 1$ among c_1, \dots, c_s are distinct. Now condition (c) of $T(\eta, \zeta)$ means that $w_1, \dots, w_{\zeta+1}$ are orthonormal to w_1, \dots, w_s for the bilinear form $u^T B v$. In particular, the orthonormality of $w_1, \dots, w_{\zeta+1}$ to $w_1, \dots, w_{\nu+1}$ (compare with (5.22)) means that the quadrature formula is exact for all polynomials of degree $\nu + \zeta$. Therefore, $p \geq \nu + \zeta + 1$ (which is the same as (5.25)) is a necessary condition for $T(\eta, \zeta)$.

To show its sufficiency, we complete $w_1, \dots, w_{\nu+1}$ to a basis of \mathbb{R}^s . The new basis vectors $\hat{w}_{\nu+2}, \dots, \hat{w}_s$ are then projected into the orthogonal complement of $\text{span}\langle w_1, \dots, w_{\zeta+1} \rangle$ with respect to $u^T B v$ by a Gram-Schmidt type orthogonalization. This yields

$$w_j = \hat{w}_j - \sum_{k=1}^{\zeta+1} (w_k^T B \hat{w}_j) w_k \quad \text{for } j = \nu+2, \dots, s. \quad \square$$

Construction of Implicit Runge-Kutta Methods

For the construction of implicit Runge-Kutta methods satisfying $B(p)$, $C(\eta)$ and $D(\zeta)$ with the help of Theorem 5.11, we first have to choose a quadrature formula of order p . The following lemma is the basic result for Gaussian integration.

Lemma 5.15. *Let c_1, \dots, c_s be real and distinct and let b_1, \dots, b_s be determined by condition $B(s)$ (i.e., the formula is “interpolatory”). Then this quadrature formula is of order $2s - k$ if and only if the polynomial $M(x) = (x - c_1)(x - c_2) \dots (x - c_s)$ is orthogonal to all polynomials of degree $\leq s - k - 1$, i.e., if and only if*

$$M(x) = C \left(P_s(x) + \alpha_1 P_{s-1}(x) + \dots + \alpha_k P_{s-k}(x) \right). \quad (5.26)$$

For a proof see Exercise 2. \square

We see from (5.26) that all quadrature formulas of order $2s - k$ can be specified in terms of k parameters $\alpha_1, \alpha_2, \dots, \alpha_k$.

Next, if the integers η and ζ satisfy $\eta + \zeta + 1 \leq 2s - k$ and $2\zeta + 1 \leq 2s - k$ (cf. (5.25)), we can compute a matrix W satisfying $T(\eta, \zeta)$ from Theorem 5.14 (or one of Lemmas 5.12 and 5.13). Finally a matrix X is chosen which satisfies (a) and (b) of Theorem 5.11. Then the Runge-Kutta method with coefficients $A = W X W^{-1}$ is of order at least $\min(\eta + \zeta + 1, 2\eta + 2)$ by Theorem 5.1.

Example 5.16. We search for all implicit Runge-Kutta methods satisfying $B(2s - 2)$, $C(s - 1)$ and $D(s - 2)$, i.e., methods which are of order at least $2s - 2$ by Theorem 5.1. As in (5.26), we put

$$M(x) = C \left(P_s(x) + \alpha_1 P_{s-1}(x) + \alpha_2 P_{s-2}(x) \right). \quad (5.27)$$

If α_2 satisfies

$$\alpha_2 < \frac{s-1}{s} \frac{\sqrt{2s+1}}{\sqrt{2s-3}},$$

then the roots of M are real and distinct (see Exercise 7). The matrix W given in (5.21) has Property $T(s-1, s-2)$ by Lemma 5.13. Finally we put

$$X = \begin{pmatrix} 1/2 & -\xi_1 & & & \\ \xi_1 & 0 & \ddots & & \\ & \ddots & \ddots & -\xi_{s-2} & \\ & & \xi_{s-2} & 0 & \beta_{s-1} \\ & & & \xi_{s-1} & \beta_s \end{pmatrix} \quad (5.28)$$

(see Theorem 5.11), and obtain with $A = W X W^{-1}$ a family of implicit Runge-Kutta methods of order $2s-2$ with the four parameters $\alpha_1, \alpha_2, \beta_s, \beta_{s-1}$.

All methods of Table 5.13 (with the exception of Lobatto IIIB) must be special cases. The corresponding parameter values are indicated in Table 5.14 (for their computation see Exercise 3). If we put $\alpha_1 = 0$ and $\alpha_2 = -\sqrt{2s+1}/\sqrt{2s-3}$ (Lobatto quadrature), we obtain the two-parameter family of Chipman (1976).

Table 5.14. Special cases of method (5.27, 5.28)

Method	α_1	α_2	β_s	β_{s-1}
Gauss	0	0	0	$-\xi_{s-1}$
Radau IA	$\sqrt{2s+1}/\sqrt{2s-1}$	0	$1/(4s-2)$	$-\xi_{s-1}$
Radau IIA	$-\sqrt{2s+1}/\sqrt{2s-1}$	0	$1/(4s-2)$	$-\xi_{s-1}$
Lobatto IIIA	0	$-\sqrt{2s+1}/\sqrt{2s-3}$	0	0
Lobatto IIIC	0	$-\sqrt{2s+1}/\sqrt{2s-3}$	$1/(2s-2)$	$-\xi_{s-1}(2s-1)/(s-1)$

Stability Function

We try to express the stability function of an implicit Runge-Kutta method in terms of the transformed Runge-Kutta matrix $X = W^{-1} A W$. From (b) and (c) of Property $T(\eta, \zeta)$ it follows that

$$W e_1 = \mathbb{1}, \quad W^T B \mathbb{1} = e_1, \quad e_1 = (1, 0, \dots, 0)^T. \quad (5.29)$$

Hence Formulas (3.2) and (3.3) become

$$R(z) = 1 + z e_1^T (I - z X)^{-1} e_1, \quad (5.30)$$

$$R(z) = \frac{\det(I - z X + z e_1 e_1^T)}{\det(I - z X)}. \quad (5.31)$$

The stability function depends only on X and not on the underlying quadrature formula. Hence, the stability function of the method of Example 5.16 depends on β_s and β_{s-1} only. Formula (5.31) becomes more symmetric (Hairer & Türke 1984) if we introduce the arithmetic mean of the matrices X and $X - e_1 e_1^T$ and define

$$Y = X - \frac{1}{2} e_1 e_1^T, \quad (5.32)$$

which is just the matrix X without the $1/2$ in the $(1, 1)$ -position.

Proposition 5.17. *For a Runge-Kutta method (3.1) let W satisfy $T(\eta, \zeta)$ for some $\eta, \zeta \geq 0$, and let Y be given by (5.32) where $X = W^{-1}AW$. The stability function then satisfies*

$$R(z) = \frac{1 + \frac{1}{2}\Psi(z)}{1 - \frac{1}{2}\Psi(z)} \quad (5.33)$$

with

$$\Psi(z) = ze_1^T(I - zY)^{-1}e_1. \quad (5.34)$$

Proof. Applying the Runge-Kutta method to the test equation (2.9) yields

$$g = \mathbb{I}y_0 + zAg, \quad y_1 = y_0 + zb^Tg.$$

With $W^{-1}g = \hat{g} = (\hat{g}_1, \dots, \hat{g}_s)^T$ this becomes

$$(I - zY)\hat{g} = e_1(y_0 + \frac{z}{2}\hat{g}_1), \quad y_1 = y_0 + z\hat{g}_1, \quad (5.35)$$

where we have used (5.29). Computing \hat{g}_1 from the first equation of (5.35) and inserting this into the second one gives the result. \square

If the Runge-Kutta method satisfies $B(2\nu + 1)$, $C(\nu)$ and $D(\nu)$ for some integer ν , then Y is given by (see Theorem 5.11)

$$Y = \begin{pmatrix} 0 & -\xi_1 & & \\ \xi_1 & \ddots & \ddots & \\ & \ddots & 0 & -\xi_\nu \\ & & \xi_\nu & \left| \begin{array}{c} Y_\nu \end{array} \right. \end{pmatrix}. \quad (5.36)$$

In this case the computation of (5.34) for the (s, s) -matrix Y can be reduced to that of the smaller $(s - \nu, s - \nu)$ -matrix Y_ν as follows:

Theorem 5.18. *If Y is given by (5.36), the function $\Psi(z)$ of (5.34) has the continued fraction representation*

$$\Psi(z) = \frac{z}{1} + \frac{\xi_1^2 z^2}{1} + \dots + \frac{\xi_{\nu-1}^2 z^2}{1} + \xi_\nu^2 z \Psi_\nu(z) \quad (5.37)$$

where $\Psi_\nu(z) = ze_1^T(I - zY_\nu)^{-1}e_1$.

Proof. Let Y_j (for $0 \leq j \leq \nu + 1$) denote the $(s - j, s - j)$ principal minors of Y , where the first j rows and columns are suppressed. Expanding the determinant of $I - zY_{j-1}$ with respect to the first row (and then the first column) gives for $j = 1, \dots, \nu$

$$\det(I - zY_{j-1}) = \det(I - zY_j) + \xi_j^2 z^2 \det(I - zY_{j+1}). \quad (5.38)$$

By Cramer's rule, the functions $\Psi_j(z)$ can also be written as

$$\Psi_j(z) = ze_1^T(I - zY_j)^{-1}e_1 = z \frac{\det(I - zY_{j+1})}{\det(I - zY_j)}. \quad (5.39)$$

Dividing (5.38) by $\det(I - zY_j)$ yields

$$\Psi_{j-1}(z) = \frac{z}{1 + \xi_j^2 z \Psi_j(z)}. \quad (5.40)$$

A repeated use of (5.40) gives (5.37) since $\Psi(z) = \Psi_0(z)$. \square

We are thus naturally led to continued fraction expansions, a technique which was historically the earliest one. Birkhoff & Varga (1965) used it in their proof of the A -stability of the diagonal Padé approximations. Later, Ehle (1969, 1973) tried to extend "Varga's proof" to verify the A -stability of the first and second subdiagonals of the Padé table ("This was unsuccessful because the resulting continued fraction expansions were not easily related to one another."). Therefore, Ehle (1973), Ehle & Picel (1975), proved A -stability results for the first and second subdiagonal and some generalizations by a completely different method. The following study of A -stability (see Butcher 1977, Hairer 1982, Hairer & Türke 1984) combines the above continued fraction expansion with properties of positive functions.

Positive Functions

Many stability conditions for numerical methods can be expressed in the form that some associated function is positive.

(G. Dahlquist 1978)

A -stability of an implicit Runge-Kutta method is defined by the property

$$|R(z)| < 1 \quad \text{for } \operatorname{Re} z < 0. \quad (5.41)$$

Since the transformation $(1 + \zeta)/(1 - \zeta)$ occurring in (5.33) maps the negative half-plane onto the open unit disc, (5.41) is equivalent to

$$\operatorname{Re} \Psi(z) < 0 \quad \text{for } \operatorname{Re} z < 0. \quad (5.42)$$

This condition means that $-\Psi(-z)$ is a positive function; for rational functions the concept of positivity can be defined as follows:

Definition 5.19. A rational function $f(z)$ is called *positive* if

$$\operatorname{Re} f(z) > 0 \quad \text{for } \operatorname{Re} z > 0.$$

A nice survey on the relevance of positive functions to numerical analysis is given by Dahlquist (1978). The following lemmas collect some properties of positive functions.

Lemma 5.20. *Let $f(z)$ and $g(z)$ be positive functions. Then we have*

- a) $\alpha f(z) + \beta g(z)$ is positive, if $\alpha > 0$ and $\beta \geq 0$;
- b) $1/f(z)$ is positive;
- c) $f(g(z))$ is positive. □

Observe that the poles of a positive function cannot lie in the positive half-plane, but poles on the imaginary axis are possible, e.g., the function $1/z$ is positive.

Lemma 5.21. *Suppose that*

$$f(z) = \frac{c}{z} + g(z) \quad \text{with} \quad g(z) = \mathcal{O}(1) \quad \text{for } z \rightarrow 0,$$

and $g(z) \not\equiv 0$. Then $f(z)$ is positive if and only if $c \geq 0$ and $g(z)$ is positive.

Proof. The “if-part” follows from Lemma 5.20. Suppose now that $f(z)$ is positive. The constant c has to be non-negative, since for small positive values of z we have $\operatorname{Re} f(z) > 0$. On the imaginary axis we have (apart from poles) $\operatorname{Re} g(iy) = \operatorname{Re} f(iy) \geq 0$ or more precisely

$$\liminf_{z \rightarrow iy, \operatorname{Re} z > 0} \operatorname{Re} g(z) \geq 0 \quad \text{for } y \in \mathbb{R}.$$

The maximum principle for harmonic functions then implies that either $g(z) \equiv 0$ or $g(z)$ is positive. □

A consequence of this lemma is the following characterization of A -stability.

Theorem 5.22. *Consider a Runge-Kutta method whose stability function is given by (5.33) with Y as in (5.36). It is A -stable if and only if*

$$\operatorname{Re} \Psi_\nu(z) < 0 \quad \text{for } \operatorname{Re} z < 0 \tag{5.43}$$

where $\Psi_\nu(z) = ze_1^T(I - zY_\nu)^{-1}e_1$ as in (5.37).

Proof. We consider the submatrices Y_j of Y and the functions $\Psi_j(z)$ of (5.39). As we prefer to work with positive functions we put

$$\chi_j(z) = -\Psi_j(-z) = ze_1^T(I + zY_\nu)^{-1}e_1. \tag{5.44}$$

By (5.42), A -stability is equivalent to the positivity of $\chi_0(z)$ and condition (5.43) means that $\chi_\nu(z)$ is a positive function. Relation (5.40) becomes

$$(\chi_{j-1}(z))^{-1} = \frac{1}{z} + \xi_j^2 \chi_j(z).$$

Since all $\chi_j(z)$ are bounded near the origin and do not vanish identically (see (5.44)), it follows from Lemma 5.21 that $\chi_j(z)$ is a positive function iff $\chi_{j-1}(z)$ is positive. This proves the theorem. □

Example 5.23. For the Runge-Kutta method of Example 5.16 with X given by (5.28) we have

$$\Psi_{s-2}(z) = \frac{z(1 - \beta_s z)}{1 - \beta_s z - \xi_{s-1} \beta_{s-1} z^2}.$$

Since

$$\left(\Psi_{s-2}(z)\right)^{-1} = \frac{1}{z} - \xi_{s-1} \beta_{s-1} \frac{z}{1 - \beta_s z}$$

it follows from Lemma 5.21 and Theorem 5.22 that the method is A -stable iff

$$\beta_{s-1} = 0 \quad \text{or} \quad (\beta_{s-1} < 0 \text{ and } \beta_s \geq 0). \quad (5.45)$$

Comparing this result with Tables 5.14 and 5.13 leads to a second proof for the A -stability of the diagonal and the first two subdiagonal Padé approximations for e^z (see Theorem 4.12).

Example 5.24 (Construction of all A -stable Runge-Kutta methods satisfying $B(2s-4)$, $C(s-2)$ and $D(s-3)$). We take a quadrature formula of order $2s-4$ and construct, by Theorem 5.14, a matrix W satisfying Property $T(s-2, s-3)$. The Runge-Kutta matrix A is then of the form

$$A = W(Y + \frac{1}{2}e_1 e_1^T)W^{-1}$$

with Y given by (5.36), $\nu = s-3$ and

$$Y_{s-3} = \begin{pmatrix} 0 & \gamma_{s-2} & \beta_{s-2} \\ \xi_{s-2} & \gamma_{s-1} & \beta_{s-1} \\ 0 & \gamma_s & \beta_s \end{pmatrix}.$$

For the study of A -stability we have to compute $\Psi_{s-3}(z)$ from (5.39). Expanding $\det(I - zY_{s-3})$ with respect to its first column we obtain

$$\left(\Psi_{s-3}(z)\right)^{-1} = \frac{1}{z} + \frac{z\xi_{s-2}(g_0 - g_1 z)}{1 - f_1 z + f_2 z^2}$$

where

$$\begin{aligned} f_1 &= \beta_s + \gamma_{s-1}, & f_2 &= \beta_s \gamma_{s-1} - \beta_{s-1} \gamma_s, \\ g_0 &= -\gamma_{s-2}, & g_1 &= -\beta_s \gamma_{s-2} + \beta_{s-2} \gamma_s. \end{aligned} \quad (5.46)$$

By Lemma 5.21 and Theorem 5.22 we have A -stability iff either $g_0 = g_1 = 0$ or

$$\frac{z(g_0 + g_1 z)}{1 + f_1 z + f_2 z^2} \quad (5.47)$$

is a positive function, which is equivalent to (see Exercise 4b)

$$g_0 > 0, \quad g_1 \geq 0, \quad f_2 \geq 0, \quad g_0 f_1 - g_1 \geq 0. \quad (5.48)$$

A similar characterization of A -stable Runge-Kutta methods of order $2s-4$ is given in Wanner (1980).

Exercises

1. Verify the integration formulas (5.10) for the shifted Legendre polynomials.

Hint. By orthogonality $\int_0^x P_k(t)dt$ must be a linear combination of P_{k+1} , P_k and P_{k-1} only. The coefficient of P_k vanishes by symmetry. For the rest just look at the coefficients of x^{k+1} and x^{k-1} .

2. Give a proof of Lemma 5.15.

Hint (Jacobi 1826). If $f(x)$ is a polynomial of degree $2s - k - 1$, and $r(x)$ the interpolation polynomial of degree $s - 1$, then $f(x) = q(x)M(x) + r(x)$, where $\deg q(x) \leq s - k - 1$.

3. Let $R(z)$ be the stability function of the Runge-Kutta method of Example 5.16.

a) The degree of its denominator is $\leq s - 1$ iff $\beta_s = \beta_{s-1}\xi_{s-1}2(2s - 3)$.

Hint. Use Formula (5.31) and the fact that $\det(I - zX_G)$ is the denominator of the diagonal Padé approximation.

b) The degree of the numerator of $R(z)$ is $\leq s - 1$ iff

$$\beta_s = -\beta_{s-1}\xi_{s-1}2(2s - 3). \quad (5.49)$$

c) The degree of the numerator of $R(z)$ is $\leq s - 2$ iff in addition to (5.49) it holds $\beta_s = 1/(2s - 2)$.

d) Verify the entries of Table 5.14.

4. a) The function

$$s(z) = \frac{\alpha + \beta z}{\gamma + \delta z}$$

with $\gamma > 0$ satisfies $\operatorname{Re} s(z) \geq 0$ for $\operatorname{Re} z > 0$ iff $\alpha \geq 0$, $\beta \geq 0$ and $\delta \geq 0$.

b) Use the identity (for $g_0 > 0$)

$$\frac{1 + f_1 z + f_2 z^2}{z(g_0 + g_1 z)} - \frac{1}{zg_0} = \frac{(f_1 - g_1/g_0) + f_2 z}{g_0 + g_1 z}$$

to verify that the function given in (5.47) is positive iff (5.48) holds.

5. Suppose that

$$f(z) = cz + g(z) \quad \text{with} \quad g(z) = \mathcal{O}(1) \quad \text{for } z \rightarrow \infty$$

and $g(z) \not\equiv 0$. Using the transformation $z \rightarrow 1/z$ in Lemma 5.21, show that $f(z)$ is a positive function, if and only if $c \geq 0$ and $g(z)$ is positive.

6. Give an alternative proof of the Routh criterion (Theorem 13.4 of Chapter I):
All zeros of the real polynomial

$$p(z) = a_0 z^n + a_1 z^{n-1} + \dots + a_n \quad (a_0 > 0)$$

lie in the negative half-plane $\operatorname{Re} z < 0$ if and only if

$$c_{i0} > 0 \quad \text{for } i = 0, 1, \dots, n.$$

The c_{ij} are the coefficients of the polynomials

$$p_i(z) = c_{i0}z^{n-i} + c_{i1}z^{n-i-2} + c_{i2}z^{n-i-4} + \dots$$

where

$$\begin{aligned} p_0(z) &= a_0z^n + a_2z^{n-2} + \dots, & \text{i.e., } c_{0j} &= a_{2j} \\ p_1(z) &= a_1z^{n-1} + a_3z^{n-3} + \dots, & \text{i.e., } c_{1j} &= a_{2j+1}. \end{aligned}$$

and

$$p_{i+1}(z) = c_{i0}p_{i-1}(z) - c_{i-1,0}zp_i(z), \quad i = 1, \dots, n-1. \quad (5.50)$$

Hint. By the maximum principle for harmonic functions the condition “ $p(z) \neq 0$ for $\operatorname{Re} z \geq 0$ ” is equivalent to “ $|p(-z)/p(z)| < 1$ for $\operatorname{Re} z > 0$ ” and the condition that $p_0(z)$ and $p_1(z)$ are irreducible. Using the transformation (5.33) this becomes equivalent to the positivity of $p_0(z)/p_1(z)$. Now divide (5.50) by $c_{i-1,0}p_i(z)$ and use Exercise 5 recursively.

7. Show that

$$\alpha_2 < \frac{s-1}{s} \frac{\sqrt{2s+1}}{\sqrt{2s-3}} \quad (5.51)$$

is a sufficient condition for $M(x) = P_s(x) + \alpha_1 P_{s-1}(x) + \alpha_2 P_{s-2}(x)$ to have real and pairwise distinct roots.

Hint. (See “Lemma 18” of Nørsett & Wanner 1981). Consider the set D of all pairs (α_1, α_2) for which the roots c_i of $M(x)$ are real and distinct, and the corresponding interpolatory quadrature formula has positive b_i . Verify that $(0, 0) \in D$, and show that for $(\alpha_1, \alpha_2) \in \partial D$ either one b_i becomes zero or two c_i coalesce but the quadrature formula remains of order $2s-2$. Therefore it must be the Gaussian formula with $s-1$ nodes of order $2s-2$ and we must have

$$P_s(x) + \alpha_1 P_{s-1}(x) + \alpha_2 P_{s-2}(x) = c(x-\beta)P_{s-1}(x). \quad (5.52)$$

Now use the three-term recursion formula

$$s\xi_s P_s(x) = (x-1/2)P_{s-1}(x) - (s-1)\xi_{s-1}P_{s-2}(x) \quad (5.53)$$

(Abramowitz & Stegun p. 782, modified) to eliminate xP_{s-1} on the right of (5.52). Then obtain by comparing the coefficients of P_s , P_{s-1} and P_{s-2}

$$c = \frac{1}{s\xi_s} \quad \alpha_1 = \frac{1}{s\xi_s} \left(\frac{1}{2} - \beta \right), \quad \alpha_2 = \frac{s-1}{s} \frac{\sqrt{2s+1}}{\sqrt{2s-3}}. \quad (5.54)$$

If β is one of the roots of P_{s-1} , then (5.52) has a double root and the estimate (5.51) for α_2 is optimal.

IV.6 Diagonally Implicit RK Methods

... they called their methods “diagonally implicit”, a term which is reserved here for the special case where all diagonal entries are equal ...
(R. Alexander 1977)

We continue to quote from this nice paper: “To integrate a system of n differential equations, an implicit method with a full $s \times s$ matrix requires the solution of ns simultaneous implicit (in general nonlinear) equations in each time step (...) One way to circumvent this difficulty is to use a lower triangular matrix (a_{ij}) (i.e., a matrix with $a_{ij} = 0$ for $i < j$); the equations may then be solved in s successive stages with only an n -dimensional system to be solved at each stage”. In accordance with many authors, and in disaccordance with others (see above), we call such a method *diagonally implicit* (DIRK).

“In solving the n -dimensional systems by Newton-type iterations one solves linear systems at each stage with a coefficient matrix of the form $I - ha_{ii}\partial f/\partial y$. If all a_{ii} are equal one may hope to use repeatedly the stored LU-factorization of a single such matrix”. When we want to emphasize this additional property for a DIRK method, we shall call it a *singly diagonally implicit* (SDIRK) method.

It is a curious coincidence that in the early seventies at least four theses dedicated a large part of their research to DIRK and SDIRK methods, very often having in mind their usefulness for the treatment of partial differential equations (R. Alt 1971, M. Crouzeix 1975, A. Kurdi 1974, S.P. Nørsett 1974). The classical paper on the subject is Alexander (1977).

Order Conditions

The traditional problem of choosing the coefficients leads to a nonlinear algebraic jungle, to which civilization and order were brought in the pioneering work of J.C. Butcher, further refined in the Thesis of M. Crouzeix.
(R. Alexander 1977)

We want to make the “jungle” still a little more civilized by the following idea: consider a SDIRK scheme

c_1	γ			
c_2	a_{21}	γ		
\vdots	\vdots	\vdots	\ddots	
c_s	a_{s1}	a_{s2}	\dots	γ
	b_1	b_2	\dots	b_s

with s stages. The order conditions (see Vol. I, Sect. II.2) consist of sums such as

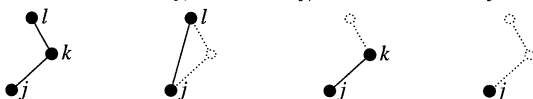
$$\sum_{j,k,l} b_j a_{jk} a_{kl} = \frac{1}{6}. \quad (6.1)$$

Because there are now more non-zero entries in the matrix A than for explicit methods, this sum contains far more terms as it did before. The trick is to transfer all expressions containing a γ to the right-hand side of (6.1). The resulting sum, denoted by \sum' , is then only built upon the subdiagonal entries as for explicit Runge-Kutta methods. The right-hand side becomes (for this example)

$$\sum_{j,k,l}' b_j a_{jk} a_{kl} = \sum_{j,k,l} b_j (a_{jk} - \gamma \delta_{jk})(a_{kl} - \gamma \delta_{kl}) \quad (6.1')$$

where δ_{jk} denotes the Kronecker delta. Multiplying out we obtain

$$\sum_{j,k,l}' b_j a_{jk} a_{kl} = \sum_{j,k,l} b_j a_{jk} a_{kl} - \gamma \left(\sum_{j,l} b_j a_{jl} + \sum_{j,k} b_j a_{jk} \right) + \gamma^2 \sum_j b_j.$$



For all sums on the right we insert order conditions (e.g. from Theorem 2.1 of Sect. II.2) and obtain

$$\sum_{j,k,l}' b_j a_{jk} a_{kl} = \frac{1}{6} - \gamma + \gamma^2. \quad (6.1'')$$

The general rule is that there appears an alternating polynomial in γ whose coefficients are sums of $1/\gamma(u)$, where u runs through all trees which are obtained by “short-circuiting” one, two, three, etc. vertices of t (with exception of the root). The conditions for order 4 obtained in this way are summarized in Table 6.1. For $s = 2$, $p = 3$ and $s = 3$, $p = 4$ these simplified conditions have only very few non-zero terms and the equations become especially simple to solve (see Exercise 1).

Stiffly Accurate SDIRK Methods


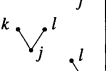
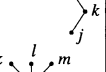
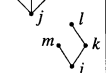
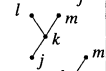
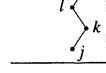
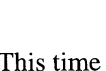
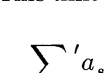
Our main interest here lies in methods satisfying

$$a_{sj} = b_j \quad \text{for } j = 1, \dots, s, \quad (6.2)$$

i.e., in methods for which the numerical solution y_1 is identical to the last internal stage. A first consequence of this property is that $R(\infty) = 0$ (see Proposition 3.8). The order conditions for such methods can, instead of (6.1''), be simplified still further: consider again the example (6.1), which can now be written as

$$\sum_{j,k,l} a_{sj} a_{jk} a_{kl} = \frac{1}{6}.$$

Table 6.1. Order conditions for SDIRK methods

t	$q(t)$	previous conditions	simplified conditions
	1	$\sum b_j = 1$	$\sum b_j = 1$
	2	$\sum b_j a_{jk} = \frac{1}{2}$	$\sum' b_j a_{jk} = \frac{1}{2} - \gamma$
	3	$\sum b_j a_{jk} a_{jl} = \frac{1}{3}$	$\sum' b_j a_{jk} a_{jl} = \frac{1}{3} - \gamma + \gamma^2$
	3	$\sum b_j a_{jk} a_{kl} = \frac{1}{6}$	$\sum' b_j a_{jk} a_{kl} = \frac{1}{6} - \gamma + \gamma^2$
	4	$\sum b_j a_{jk} a_{jl} a_{jm} = \frac{1}{4}$	$\sum' b_j a_{jk} a_{jl} a_{jm} = \frac{1}{4} - \gamma + \frac{3}{2}\gamma^2 - \gamma^3$
	4	$\sum b_j a_{jk} a_{kl} a_{jm} = \frac{1}{8}$	$\sum' b_j a_{jk} a_{kl} a_{jm} = \frac{1}{8} - \frac{5}{6}\gamma + \frac{3}{2}\gamma^2 - \gamma^3$
	4	$\sum b_j a_{jk} a_{kl} a_{km} = \frac{1}{12}$	$\sum' b_j a_{jk} a_{kl} a_{km} = \frac{1}{12} - \frac{2}{3}\gamma + \frac{3}{2}\gamma^2 - \gamma^3$
	4	$\sum b_j a_{jk} a_{kl} a_{lm} = \frac{1}{24}$	$\sum' b_j a_{jk} a_{kl} a_{lm} = \frac{1}{24} - \frac{1}{2}\gamma + \frac{3}{2}\gamma^2 - \gamma^3$

This time we have, instead of (6.1')

$$\begin{aligned}
 \sum_{j,k,l}' a_{sj} a_{jk} a_{kl} &= \sum_{j,k,l} (a_{sj} - \gamma \delta_{sj})(a_{jk} - \gamma \delta_{jk})(a_{kl} - \gamma \delta_{kl}) \\
 &= \sum_{j,k,l} a_{sj} a_{jk} a_{kl} - \gamma \left(\sum_{j,k} a_{sj} a_{jk} + \sum_{j,l} a_{sj} a_{jl} + \sum_{k,l} a_{sk} a_{kl} \right) \\
 &\quad + \gamma^2 \left(\sum_j a_{sj} + \sum_k a_{sk} + \sum_l a_{sl} \right) - \gamma^3 \cdot 1.
 \end{aligned}$$

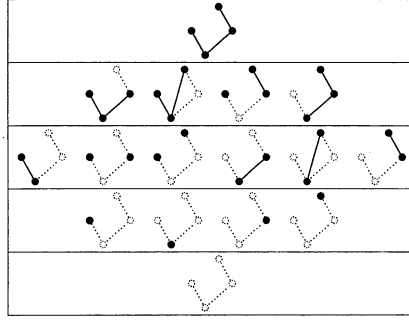
Again inserting known order conditions, we now obtain

$$\sum_{j,k,l}' a_{sj} a_{jk} a_{kl} = \frac{1}{6} - \frac{3}{2}\gamma + 3\gamma^2 - \gamma^3. \quad (6.1''')$$

The general rule is similar to the one above: the difference is that *all* vertices (including the root) are now available for being short-circuited. Another example, for the tree t_{42} , is sketched in Fig. 6.1 and leads to the following right-hand side:

$$\begin{aligned}
 &\frac{1}{8} - \gamma \left(\frac{1}{3} + \frac{1}{3} + 1 \cdot \frac{1}{2} + \frac{1}{6} \right) + \gamma^2 \left(\frac{1}{2} + 1 \cdot 1 + 1 \cdot 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} \right) \\
 &\quad - \gamma^3 (1 + 1 + 1 + 1) + \gamma^4 = \frac{1}{8} - \frac{4}{3}\gamma + 4\gamma^2 - 4\gamma^3 + \gamma^4.
 \end{aligned}$$

The order conditions obtained in this manner are displayed in Table 6.2 for all trees of order ≤ 4 . The expressions \sum' are written explicitly for the SDIRK method

**Fig. 6.1.** Short-circuiting tree t_{42}

(6.3) with $s = 5$ satisfying condition (6.2)

γ							
a_{21}	γ						$c'_2 = a_{21}$
a_{31}	a_{32}	γ					$c'_3 = a_{31} + a_{32}$
a_{41}	a_{42}	a_{43}	γ				$c'_4 = a_{41} + a_{42} + a_{43}$
b_1	b_2	b_3	b_4	γ			
b_1	b_2	b_3	b_4	γ			

(6.3)

Observe that they become very similar to those of Formulas (1.11) in Sect. II.1.

Table 6.2. Order conditions for method (6.3)

	$\sum' a_{sj} = b_1 + b_2 + b_3 + b_4 = p_1$	(6.4;1)
	$\sum' a_{sj} a_{jk} = b_2 c'_2 + b_3 c'_3 + b_4 c'_4 = p_2$	(6.4;2)
	$\sum' a_{sj} a_{jk} a_{jl} = b_2 c'^2_2 + b_3 c'^2_3 + b_4 c'^2_4 = p_3$	(6.4;3)
	$\sum' a_{sj} a_{jk} a_{kl} = b_3 a_{32} c'_2 + b_4 (a_{42} c'_2 + a_{43} c'_3) = p_4$	(6.4;4)
	$\sum' a_{sj} a_{jk} a_{jl} a_{jm} = b_2 c'^3_2 + b_3 c'^3_3 + b_4 c'^3_4 = p_5$	(6.4;5)
	$\sum' a_{sj} a_{jk} a_{jl} a_{lm} = b_3 c'_3 a_{32} c'_2 + b_4 c'_4 (a_{42} c'_2 + a_{43} c'_3) = p_6$	(6.4;6)
	$\sum' a_{sj} a_{jk} a_{kl} a_{km} = b_3 a_{32} c'^2_2 + b_4 (a_{42} c'^2_2 + a_{43} c'^2_3) = p_7$	(6.4;7)
	$\sum' a_{sj} a_{jk} a_{kl} a_{lm} = b_4 a_{43} a_{32} c'_2 = p_8$	(6.4;8)

$p_1 = 1 - \gamma$	$p_5 = \frac{1}{4} - 2\gamma + \frac{9}{2}\gamma^2 - 4\gamma^3 + \gamma^4$
$p_2 = \frac{1}{2} - 2\gamma + \gamma^2$	$p_6 = \frac{1}{8} - \frac{4}{3}\gamma + 4\gamma^2 - 4\gamma^3 + \gamma^4$
$p_3 = \frac{1}{3} - 2\gamma + 3\gamma^2 - \gamma^3$	$p_7 = \frac{1}{12} - \gamma + \frac{7}{2}\gamma^2 - 4\gamma^3 + \gamma^4$
$p_4 = \frac{1}{6} - \frac{3}{2}\gamma + 3\gamma^2 - \gamma^3$	$p_8 = \frac{1}{24} - \frac{2}{3}\gamma + 3\gamma^2 - 4\gamma^3 + \gamma^4$

Solution of Equations (6.4). By clever elimination from equations (6.4;4) and (6.4;6) as well as (6.4;4) and (6.4;7) we obtain

$$\begin{aligned} b_3 a_{32} c'_2 (c'_4 - c'_3) &= c'_4 p_4 - p_6 \\ b_4 c'_3 a_{43} (c'_2 - c'_3) &= c'_2 p_4 - p_7. \end{aligned} \quad (6.5)$$

Multiplying these two equations and using (6.4;8) gives

$$p_8 b_3 (c'_4 - c'_3) (c'_2 - c'_3) c'_3 = (c'_4 p_4 - p_6) (c'_2 p_4 - p_7).$$

We now compute b_2, b_3, b_4 from (6.4;2), (6.4;3), (6.4;5). This gives

$$b_3 = (-p_2 c'_2 c'_4 + p_3 (c'_4 + c'_2) - p_5) / (c'_3 (c'_3 - c'_2) (c'_4 - c'_3)) \quad (6.6)$$

and b_2 as well as b_4 by cyclic permutation. Comparing the last two equations leads to

$$c'_4 = \frac{p_8 p_3 c'_2 - p_8 p_5 - c'_2 p_6 p_4 + p_6 p_7}{p_8 p_2 c'_2 - p_8 p_3 - c'_2 p_4 p_4 + p_4 p_7}. \quad (6.7)$$

We now choose γ, c'_2 and c'_3 as free parameters. Then c'_4 is obtained from (6.7); b_2, b_3, b_4 from (6.6), b_1 from (6.4;1), a_{32} and a_{43} from (6.5), a_{42} from (6.4;4), and finally a_{21}, a_{31}, a_{41} from (6.3).

Embedded 3rd order formula: As proposed by Cash (1979), we can append to the above formula a third order expression

$$\hat{y}_1 = y_0 + h \sum_{i=1}^4 \hat{b}_i k_i$$

(thus by omitting the term $b_5 = \gamma$) for the sake of step size control. The coefficients $\hat{b}_1, \dots, \hat{b}_4$ are simply obtained by solving the first 4 equations of Table 6.1 (linear system). *Continuous* embedded 3rd order formulas can be obtained in this way too (see Theorem 6.1 of Sect. II.6)

$$y(x_0 + \theta h) \approx y_0 + h \sum_{i=1}^4 b_i(\theta) k_i.$$

The coefficients $b_1(\theta), \dots, b_4(\theta)$ are obtained by solving the first 4 (simplified) conditions of Table 6.1, with the right-hand sides replaced by

$$\theta, \quad \frac{\theta^2}{2} - \gamma\theta, \quad \frac{\theta^3}{3} - \gamma\theta^2 + \gamma^2\theta, \quad \frac{\theta^3}{6} - \gamma\theta^2 + \gamma^2\theta,$$

respectively. The continuous solution obtained in this way becomes \hat{y}_1 for $\theta = 1$ instead of the 4-th order solution y_1 . The global continuous solution would therefore be discontinuous. In order to avoid this discontinuity, we add $b_5(\theta)$ and include the fifth equation from Table 6.1 with right-hand side

$$\frac{\theta^4}{4} - \gamma\theta^3 + \frac{3\gamma^2\theta^2}{2} - \gamma^3\theta.$$

The Stability Function

By Formula (3.3), the stability function $R(z)$ for a DIRK method is of the form

$$R(z) = \frac{P(z)}{(1 - a_{11}z)(1 - a_{22}z) \dots (1 - a_{ss}z)}, \quad (6.8)$$

because the determinant of a triangular matrix is the product of its diagonal entries. The numerator $P(z)$ is a polynomial of degree s at most. If the method is of order $p \geq s$, this polynomial is uniquely determined by Formula (3.18). It is simply obtained from the first terms of the power series for $(1 - a_{11}z) \dots (1 - a_{ss}z) \cdot e^z$.

For SDIRK methods, with $a_{11} = \dots = a_{ss} = \gamma$, we obtain (see also Formula (3.18) with $q_j = (-\gamma)^j \binom{s}{j}$)

$$R(z) = \frac{P(z)}{(1 - \gamma z)^s}, \quad P(z) = (-1)^s \sum_{j=0}^s L_s^{(s-j)}\left(\frac{1}{\gamma}\right) (\gamma z)^j \quad (6.9)$$

with error constant

$$C = \frac{\gamma^s (-1)^{s+1}}{s+1} L_{s+1}^{(1)}\left(\frac{1}{\gamma}\right) \quad (6.10)$$

where

$$L_s(x) = \sum_{j=0}^s (-1)^j \binom{s}{j} \frac{x^j}{j!} \quad (6.11)$$

is the s -degree Laguerre polynomial. $L_s^{(k)}(x)$ denotes its k -th derivative. Since the function (6.9) is analytic in \mathbb{C}^- for $\gamma > 0$, A -stability is equivalent to

$$E(y) = Q(iy)Q(-iy) - P(iy)P(-iy) \geq 0 \quad \text{for all } y \quad (6.12)$$

(see (3.8)). This is an even polynomial of degree $2s$ (in general) and subdegree $2j$ where $j = [(p+2)/2]$ (see Proposition 3.4). We therefore define the polynomial $F(x)$ by

$$F(y^2) = E(y)/y^{2j} \quad j = [(p+2)/2].$$

and check the condition $F(x) \geq 0$ for $x \geq 0$ using Sturm sequences. We display the results obtained (similar to Burrage 1978) in Table 6.3.

For completeness, we give the following explicit formulas for $E(y)$.

$s = 1$; $p = 1$:

$$E = y^2(2\gamma - 1)$$

$s = 2$; $p = 2$:

$$E = y^4 \left(-\frac{1}{4} + 2\gamma - 5\gamma^2 + 4\gamma^3 \right) = y^4 (2\gamma - 1)^2 \left(\gamma - \frac{1}{4} \right)$$

$s = 3$; $p = 3$:

$$E = y^4 \left(\frac{1}{12} - \gamma + 3\gamma^2 - 2\gamma^3 \right) + y^6 \left(-\frac{1}{36} + \frac{\gamma}{2} - \frac{13\gamma^2}{4} + \frac{28\gamma^3}{3} - 12\gamma^4 + 6\gamma^5 \right)$$

Table 6.3. A -stability of (6.9), order $p \geq s$

s	A -stability	A -stability and $p = s + 1$
1	$1/2 \leq \gamma < \infty$	$1/2$
2	$1/4 \leq \gamma < \infty$	$(3 + \sqrt{3})/6$
3	$1/3 \leq \gamma \leq 1.06857902$	1.06857902
4	$0.39433757 \leq \gamma \leq 1.28057976$	—
5	$\begin{cases} 0.24650519 \leq \gamma \leq 0.36180340 \\ 0.42078251 \leq \gamma \leq 0.47326839 \end{cases}$	0.47326839
6	$0.28406464 \leq \gamma \leq 0.54090688$	—
7	—	—
8	$0.21704974 \leq \gamma \leq 0.26471425$	—

$s = 4; p = 4 :$

$$E = y^6 \left(\frac{1}{72} - \frac{\gamma}{3} + \frac{17\gamma^2}{6} - \frac{32\gamma^3}{3} + 17\gamma^4 - 8\gamma^5 \right) \\ + y^8 \left(-\frac{1}{576} + \frac{\gamma}{18} - \frac{25\gamma^2}{36} + \frac{13\gamma^3}{3} - \frac{173\gamma^4}{12} + \frac{76\gamma^5}{3} - 22\gamma^6 + 8\gamma^7 \right).$$

A -stability means here that all coefficients must be non-negative. A general formula is as follows.

Lemma 6.1. *The E -polynomial for (6.8) with $a_{11} = \dots = a_{ss} = \gamma$ and $p \geq s$ satisfies*

$$E(y) = \left(1 - L_s \left(\frac{1}{\gamma} \right)^2 \right) (\gamma y)^{2s} \\ - 2 \sum_{j=[(p+2)/2]}^{s-1} (-1)^{s+j} (\gamma y)^{2j} \int_0^{1/\gamma} L_s(x) L_s^{(2s+1-2j)}(x) dx. \quad (6.13)$$

Proof. Inserting Formula (6.9) into the definition of $E(y)$

$$E(y) = (1 + \gamma^2 y^2)^s - P(iy)P(-iy) \\ = (1 + \gamma^2 y^2)^s - \sum_k \sum_l L_s^{(s-k)} \left(\frac{1}{\gamma} \right) L_s^{(s-l)} \left(\frac{1}{\gamma} \right) (\gamma iy)^{k+l} (-1)^l$$

and using integration by parts for the verification of

$$2 \int_0^\alpha L_s(x) L_s^{(2s+1-2j)}(x) dx = (-1)^s \sum_{k+l=2j} (-1)^l L_s^{(s-k)}(x) L_s^{(s-l)}(x) \Big|_0^\alpha$$

one obtains the result, since

$$\sum_{k+l=2j} (-1)^l L_s^{(s-k)}(0) L_s^{(s-l)}(0) = (-1)^j \binom{s}{j}. \quad \square$$

Multiple Real-Pole Approximations with $R(\infty) = 0$

For methods satisfying (6.2) we have $R(\infty) = 0$. Therefore the highest coefficient of $P(z)$ in (6.9) is zero. If the order of the method is known to be $p \geq s - 1$, the remaining coefficients of $P(z)$ are still uniquely determined by γ and we have

$$P(z) = (-1)^s \sum_{j=0}^{s-1} L_s^{(s-j)} \left(\frac{1}{\gamma} \right) (\gamma z)^j \quad (6.14)$$

with error constant

$$C = (-1)^s L_s \left(\frac{1}{\gamma} \right) \gamma^s. \quad (6.15)$$

The first polynomials $E(y)$ of (6.12) are now:

$s = 2, p = 1:$

$$E = y^2(-1 + 4\gamma - 2\gamma^2) + y^4\gamma^4$$

$s = 3, p = 2:$

$$E = y^4 \left(-\frac{1}{4} + 3\gamma - 12\gamma^2 + 18\gamma^3 - 6\gamma^4 \right) + y^6\gamma^6$$

$s = 4, p = 3:$

$$E = y^4 \left(\frac{1}{12} - \frac{4\gamma}{3} + 6\gamma^2 - 8\gamma^3 + 2\gamma^4 \right) + y^6 \left(-\frac{1}{36} + \frac{2\gamma}{3} - 6\gamma^2 + \frac{76\gamma^3}{3} - 52\gamma^4 + 48\gamma^5 - 12\gamma^6 \right) + y^8\gamma^8.$$

The regions of γ for A -(and hence L -)stability are displayed in Table 6.4.

Table 6.4. L -stability of $R(z)$ with P from (6.14), order $p \geq s - 1$

s	L -stability	L -stab. and $p = s$
2	$(2 - \sqrt{2})/2 \leq \gamma \leq (2 + \sqrt{2})/2$	$\gamma = (2 \pm \sqrt{2})/2$
3	$0.18042531 \leq \gamma \leq 2.18560010$	$\gamma = 0.43586652$
4	$0.22364780 \leq \gamma \leq 0.57281606$	$\gamma = 0.57281606$
5	$0.24799464 \leq \gamma \leq 0.67604239$	$\gamma = 0.27805384$
6	$0.18391465 \leq \gamma \leq 0.33414237$	$\gamma = 0.33414237$
7	$0.20408345 \leq \gamma \leq 0.37886489$	—
8	$0.15665860 \leq \gamma \leq 0.23437316$	$\gamma = 0.23437316$

Choice of Method

We now determine the free parameters for method (6.3) with $s = 5$ and order 4. For a good choice of γ , we have displayed in Fig. 6.2 the error constant C as well as the regions for A - and $A(0)$ -stability.

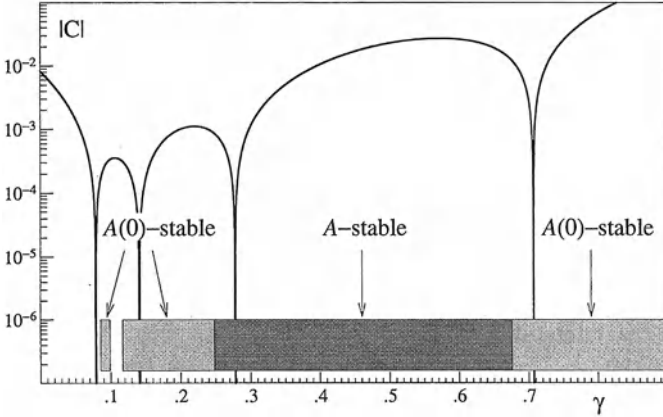


Fig. 6.2 Error constant and A -stability domain for $s = 5$, $p = 4$

This suggests that γ between 0.25 and 0.29 is a good choice. The method is then L -stable and the error constant is small. For various values of γ in this range, we determined (by a nonlinear Gauss-Newton code) c'_2 and c'_3 in order to minimize the fifth-order error terms. It turned out that

$$c'_2 = 0.5, \quad c'_3 = 0.3$$

is close to optimal. With this we coded two different choices of γ : $\gamma = 4/15 = 0.2666\dots$, which was numerically the better choice and $\gamma = 1/4$, which gave, via Formulas (6.4), (6.5), (6.6) and (6.7), especially nice rational coefficients. These latter are displayed in Table 6.5. We have included a continuous solution to this method

$$y(x_0 + \theta h) \approx y_0 + h \sum_{j=1}^5 b_j(\theta) k_j,$$

which is third order for $0 < \theta < 1$ and updates to the fourth order approximation y_1 for $\theta = 1$.

Table 6.5. *L*-stable SDIRK method of order 4

$\frac{1}{4}$	$\frac{1}{4}$				
$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$			
$\frac{11}{20}$	$\frac{17}{50}$	$-\frac{1}{25}$	$\frac{1}{4}$		
$\frac{1}{2}$	$\frac{371}{1360}$	$-\frac{137}{2720}$	$\frac{15}{544}$	$\frac{1}{4}$	
1	$\frac{25}{24}$	$-\frac{49}{48}$	$\frac{125}{16}$	$-\frac{85}{12}$	$\frac{1}{4}$
$y_1 =$	$\frac{25}{24}$	$-\frac{49}{48}$	$\frac{125}{16}$	$-\frac{85}{12}$	$\frac{1}{4}$
$\hat{y}_1 =$	$\frac{59}{48}$	$-\frac{17}{96}$	$\frac{225}{32}$	$-\frac{85}{12}$	0
$err =$	$-\frac{3}{16}$	$-\frac{27}{32}$	$\frac{25}{32}$	0	$\frac{1}{4}$

(6.16)

$$\begin{aligned}
 b_1(\theta) &= \frac{11}{3}\theta - \frac{463}{72}\theta^2 + \frac{217}{36}\theta^3 - \frac{20}{9}\theta^4 \\
 b_2(\theta) &= \frac{11}{2}\theta - \frac{385}{16}\theta^2 + \frac{661}{24}\theta^3 - 10\theta^4 \\
 b_3(\theta) &= -\frac{125}{18}\theta + \frac{20125}{432}\theta^2 - \frac{8875}{216}\theta^3 + \frac{250}{27}\theta^4 \\
 b_4(\theta) &= -\frac{85}{4}\theta^2 + \frac{85}{6}\theta^3 \\
 b_5(\theta) &= -\frac{11}{9}\theta + \frac{557}{108}\theta^2 - \frac{359}{54}\theta^3 + \frac{80}{27}\theta^4.
 \end{aligned}
 \tag{6.17}$$

Exercises

1. (Crouzeix & Raviart 1980). Compute the SDIRK methods (Table 6.1) for $s = 3$, $p = 4$. Obtain also (for $s = 2$, $p = 3$) once again the method of Table 7.2, Sect. II.7.

Result. The last order condition is in both cases just a polynomial in γ . Among the different solutions, the following presents an *A*-stable scheme:

γ	γ
$\frac{1}{2}$	$\frac{1}{2} - \gamma$ γ
$1 - \gamma$	2γ $1 - 4\gamma$ γ
	δ $1 - 2\delta$ δ

$$\begin{aligned}
 \gamma &= \frac{1}{\sqrt{3}} \cos\left(\frac{\pi}{18}\right) + \frac{1}{2} \\
 \delta &= \frac{1}{6(2\gamma - 1)^2}.
 \end{aligned}
 \tag{6.18}$$

2. Verify all details of Tables 6.1 and 6.2.

3. The four cases of A -stable SDIRK methods of order $p = s + 1$ indicated in Table 6.3 (right) are the *only* ones existing. This fact has not yet been *rigorously* proved, because the “proof” given in Wanner, Hairer & Nørsett (1978) uses an asymptotic formula without error estimation. Do better.
4. Cooper & Sayfy (1979) have derived many DIRK (which they call “semi-implicit”) methods of high order. Their main aim was to *minimize* the number of implicit stages and *not* to maximize stability. One of their methods is

$\frac{6-\sqrt{6}}{10}$	$\frac{6-\sqrt{6}}{10}$				
$\frac{6+9\sqrt{6}}{35}$	$\frac{-6+5\sqrt{6}}{14}$	$\frac{6-\sqrt{6}}{10}$			
1	$\frac{888+607\sqrt{6}}{2850}$	$\frac{126-161\sqrt{6}}{1425}$	$\frac{6-\sqrt{6}}{10}$		
$\frac{4-\sqrt{6}}{10}$	$\frac{3153-3082\sqrt{6}}{14250}$	$\frac{3213+1148\sqrt{6}}{28500}$	$\frac{-267+88\sqrt{6}}{500}$	$\frac{6-\sqrt{6}}{10}$	
$\frac{4+\sqrt{6}}{10}$	$\frac{-32583+14638\sqrt{6}}{71250}$	$\frac{-17199+364\sqrt{6}}{142500}$	$\frac{1329-544\sqrt{6}}{2500}$	$\frac{-96+131\sqrt{6}}{625}$	$\frac{6-\sqrt{6}}{10}$
1	0	0	$\frac{1}{9}$	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$ 0

Show that it is of order 5 and A -stable, but not L -stable.

5. It can be seen in Table 6.4 that for $s = 2, 4, 6$, and 8 the L -stability superconvergence point coincides with the right end of the A -stability interval. Explain this with the help of order star theory (Fig. 6.3.a).

Further, for $s = 7$, a superconvergence point is given by $\gamma = 0.20406693$, which misses the A -stability interval given there by less than $2 \cdot 10^{-5}$. Should the above argument also apply here and must there be a computation error somewhere? Study the corresponding order star to show that this is not the case (Fig. 6.3.b).

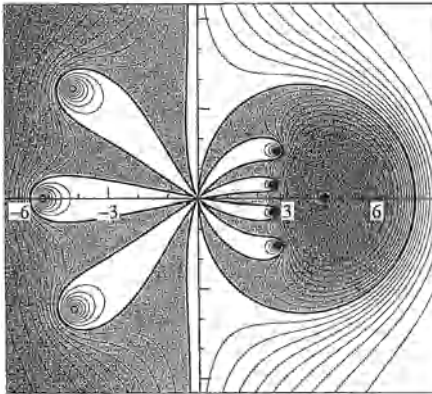


Fig. 6.3.a.
Multiple pole order star
 $s = 8$, $\gamma = 0.23437316$

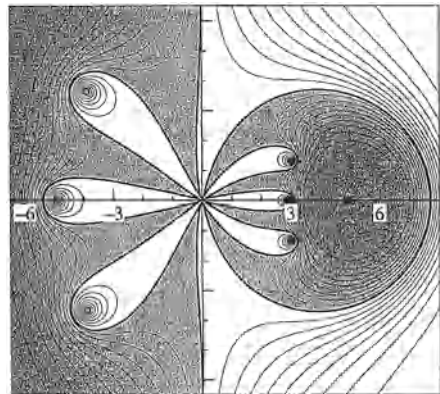


Fig. 6.3.b.
Multiple pole order star
 $s = 7$, $\gamma = 0.20406693$

IV.7 Rosenbrock-Type Methods

When the functions φ are non-linear, implicit equations can in general be solved only by iteration. This is a severe drawback, as it adds to the problem of stability, that of convergence of the iterative process. An alternative, which avoids this difficulty, is
 ... (H.H. Rosenbrock 1962/63)

... is discussed in this section. Among the methods which already give satisfactory results for stiff equations, Rosenbrock methods are the easiest to program. We shall describe their theory in this section, which will lead us to our first “stiff” code. Rosenbrock methods belong to a large class of methods which try to avoid nonlinear systems and replace them by a sequence of linear systems. We therefore call these methods *linearly implicit Runge-Kutta methods*. In the literature such methods are often called “semi-implicit” (or was it “semi-explicit”?), or “generalized” or “modified” or “adaptive” or “additive” Runge-Kutta methods.

Derivation of the Method

We start, say, with a diagonally implicit Runge-Kutta method

$$\begin{aligned} k_i &= hf \left(y_0 + \sum_{j=1}^{i-1} a_{ij} k_j + a_{ii} k_i \right) \quad i = 1, \dots, s \\ y_1 &= y_0 + \sum_{i=1}^s b_i k_i \end{aligned} \tag{7.1}$$

applied to the autonomous differential equation

$$y' = f(y). \tag{7.2}$$

The main idea is to linearize Formula (7.1). This yields

$$\begin{aligned} k_i &= hf(g_i) + hf'(g_i) a_{ii} k_i \\ g_i &= y_0 + \sum_{j=1}^{i-1} a_{ij} k_j, \end{aligned} \tag{7.3}$$

and can be interpreted as the application of *one* Newton iteration to each stage in (7.1) with starting values $k_i^{(0)} = 0$. Instead of continuing the iterations until convergence, we consider (7.3) as a new class of methods and investigate anew its order and stability properties.

Important computational advantage is obtained by replacing the Jacobians $f'(g_i)$ by $J = f'(y_0)$, so that the method requires its calculation only once (Calahan 1968). Many methods of this type and much numerical experience with them have been obtained by van der Houwen (1973), Cash (1976) and Nørsett (1975).

We gain further freedom by introducing additional linear combinations of the terms Jk_j into (7.3) (Nørsett & Wolfbrandt 1979, Kaps & Rentrop 1979). We then arrive at the following class of methods:

Definition 7.1. An s -stage Rosenbrock method is given by the formulas

$$\begin{aligned} k_i &= hf \left(y_0 + \sum_{j=1}^{i-1} \alpha_{ij} k_j \right) + hJ \sum_{j=1}^i \gamma_{ij} k_j, \quad i = 1, \dots, s \\ y_1 &= y_0 + \sum_{j=1}^s b_j k_j \end{aligned} \quad (7.4)$$

where $\alpha_{ij}, \gamma_{ij}, b_i$ are the determining coefficients and $J = f'(y_0)$.

Each stage of this method consists of a system of linear equations with unknowns k_i and with matrix $I - h\gamma_{ii}J$. Of special interest are methods for which $\gamma_{11} = \dots = \gamma_{ss} = \gamma$, so that we need only one LU-decomposition per step.

Non-autonomous problems. The equation

$$y' = f(x, y) \quad (7.2a)$$

can be converted to autonomous form by adding $x' = 1$. If method (7.4) is applied to the augmented system, the components corresponding to the x -variable can be computed explicitly and we arrive at

$$\begin{aligned} k_i &= hf \left(x_0 + \alpha_i h, y_0 + \sum_{j=1}^{i-1} \alpha_{ij} k_j \right) + \gamma_i h^2 \frac{\partial f}{\partial x}(x_0, y_0) + h \frac{\partial f}{\partial y}(x_0, y_0) \sum_{j=1}^i \gamma_{ij} k_j \\ y_1 &= y_0 + \sum_{j=1}^s b_j k_j, \end{aligned} \quad (7.4a)$$

where the additional coefficients are given by

$$\alpha_i = \sum_{j=1}^{i-1} \alpha_{ij}, \quad \gamma_i = \sum_{j=1}^i \gamma_{ij}. \quad (7.5)$$

Implicit differential equations. Suppose the problem is of the form

$$My' = f(x, y) \quad (7.2b)$$

where M is a constant matrix (nonsingular for the moment). If we formally multiply (7.2b) with M^{-1} , apply method (7.4a), and then multiply the resulting formula

with M , we obtain

$$Mk_i = hf\left(x_0 + \alpha_i h, y_0 + \sum_{j=1}^{i-1} \alpha_{ij} k_j\right) + \gamma_i h^2 \frac{\partial f}{\partial x}(x_0, y_0) + h \frac{\partial f}{\partial y}(x_0, y_0) \sum_{j=1}^i \gamma_{ij} k_j$$

$$y_1 = y_0 + \sum_{j=1}^s b_j k_j. \quad (7.4b)$$

An advantage of this formulation is that the inversion of M is avoided and that possible band-structures of the matrices M and $\partial f/\partial y$ are preserved.

Order Conditions

Conditions on the free parameters which ensure that the method is of order p , i.e., the local error satisfies

$$y(x_0 + h) - y_1 = \mathcal{O}(h^{p+1}),$$

can be obtained either by straightforward differentiation or by the use of the theorems on B -series (Sect. II.12). We follow here the first approach, since it requires only the knowledge of Sect. II.2. The second possibility is sketched in Exercise 2.

As in Sect. II.2, we write the system (7.2) in tensor notation and Method (7.4) as ¹

$$k_j^J = hf^J(g_j) + h \sum_K f_K^J(y_0) \sum_k \gamma_{jk} k_k^K$$

$$g_i^J = y_0^J + \sum_j \alpha_{ij} k_j^J, \quad (7.4')$$

$$y_1^J = y_0^J + \sum_j b_j k_j^J.$$

Again, we use Leibniz's rule (cf. (II.2.4))

$$(k_j^J)^{(q)}|_{h=0} = q(f^J(g_j))^{(q-1)}|_{h=0} + q \sum_K f_K^J(y_0) \sum_k \gamma_{jk} (k_k^K)^{(q-1)}|_{h=0} \quad (7.6)$$

and have from the chain rule (cf. Sect. II.2, (2.6;1), (2.6;2))

$$(f^J(g_j))' = \sum_K f_K^J(g_j) \cdot (g_j^K)'$$

$$(f^J(g_j))'' = \sum_{K,L} f_{KL}^J(g_j) \cdot (g_j^K)' \cdot (g_j^L)' + \sum_K f_K^J(g_j) \cdot (g_j^K)''$$

¹ In the sequel, the reader will find many k 's of different meaning; on the one hand the " k " in Formula (7.1) which goes back to Runge and Kutta, on the other hand " k " as summation index as since ever in numerical analysis. Although this looks somewhat strange in certain formulas, we prefer to retain the notation of previous sections.

etc. Inserting this into (7.6) we obtain recursively

$$(k_j^J)^{(0)}|_{h=0} = 0 \quad (7.7;0)$$

$$(k_j^J)^{(1)}|_{h=0} = f^J \quad (7.7;1)$$

$$\begin{aligned} (k_j^J)^{(2)}|_{h=0} &= 2 \sum_K f_K^J f^K \sum_k \alpha_{jk} + 2 \sum_K f_K^J f^K \sum_k \gamma_{jk} \\ &= 2 \sum_K f_K^J f^K \sum_k (\alpha_{jk} + \gamma_{jk}) \end{aligned} \quad (7.7;2)$$

$$\begin{aligned} (k_j^J)^{(3)}|_{h=0} &= 3 \sum_{K,L} f_{KL}^J f^K f^L \sum_{k,l} \alpha_{jk} \alpha_{jl} \\ &\quad + 3 \cdot 2 \sum_{K,L} f_K^J f_L^K f^L \sum_{k,l} (\alpha_{jk} + \gamma_{jk})(\alpha_{kl} + \gamma_{kl}) \end{aligned} \quad (7.7;3)$$

etc. All elementary differentials are evaluated at y_0 . Comparing the derivatives of the numerical solution ($q \geq 1$)

$$(y_1^J)^{(q)}|_{h=0} = \sum_j b_j (k_j^J)^{(q)}|_{h=0} \quad (7.8)$$

with those of the true solution (Sect. II.2, Formula (2.7;1), (2.7;2), (2.7;3)), we arrive at the following conditions for order three:

\bullet_j	$\sum b_j = 1$
\nearrow_j^k	$\sum b_j (\alpha_{jk} + \gamma_{jk}) = \frac{1}{2}$
\nwarrow_j^l	$\sum b_j \alpha_{jk} \alpha_{jl} = \frac{1}{3}$
\searrow_j^l	$\sum b_j (\alpha_{jk} + \gamma_{jk})(\alpha_{kl} + \gamma_{kl}) = \frac{1}{6}$

The only difference with the order conditions for Runge-Kutta methods is that at singly-branched vertices of the corresponding trees α_{jk} is replaced by $\alpha_{jk} + \gamma_{jk}$. In order to arrive at a general result, the formulas obtained motivate the following definition:

Definition 7.2. Let t be a labelled tree of order q with root j ; we denote by

$$\Phi_j(t) = \sum_{k,l,\dots} \varphi_{j,k,l,\dots}$$

the sum over the remaining $q-1$ indices k, l, \dots etc. The summand $\varphi_{j,k,l,\dots}$ is a product of $q-1$ factors, which are

$$\begin{aligned} &\alpha_{kl} + \gamma_{kl} && \text{if } l \text{ is the only son of } k; \\ &\alpha_{kl} && \text{if } l \text{ is a son of } k \text{ and } k \text{ has at least two sons.} \end{aligned}$$

Using the recursive representation of trees (Def. II.2.12) we have $\Phi_j(\tau) = 1$ for the only tree of order 1 and, as in (II.2.19),

$$\Phi_j(t) = \begin{cases} \sum_{k_1, \dots, k_m} \alpha_{jk_1} \dots \alpha_{jk_m} \Phi_{k_1}(t_1) \dots \Phi_{k_m}(t_m) & \text{if } t = [t_1, \dots, t_m], \\ & m \geq 2 \\ \sum_k (\alpha_{jk} + \gamma_{jk}) \Phi_k(t_1) & \text{if } t = [t_1]. \end{cases} \quad (7.9)$$

Theorem 7.3. *The derivatives of k_j^J , given by (7.4'), satisfy*

$$(k_j^J)^{(q)}|_{h=0} = \sum_{t \in LT_q} \gamma(t) \Phi_j(t) F^J(t)(y_0) \quad (7.7;q)$$

and the numerical solution y_1^J satisfies

$$(y_1^J)^{(q)}|_{h=0} = \sum_{t \in LT_q} \gamma(t) \sum_j b_j \Phi_j(t) F^J(t)(y_0), \quad (7.10)$$

where $F^J(t)$ are the elementary differentials (Definition II.2.3).

Proof. Because of (7.8) we only have to prove the first formula. This is done by induction on q and follows exactly the lines of the proof of Theorem II.2.11. We use (7.6), replace the expression $f^J(g_j)^{(q-1)}$ by Faà di Bruno's formula (Lemma II.2.8), use

$$(g_j^K)^{(\delta)} = \sum_k \alpha_{jk} (k_k^K)^{(\delta)}$$

for the derivatives of g_j and insert the induction hypothesis (7.7; δ) with $\delta \leq q-1$. This gives

$$\begin{aligned} (k_j^J)^{(q)}|_{h=0} &= q \sum_{u \in LS_q} \sum_{t_1 \in LT_{\delta_1}} \dots \sum_{t_m \in LT_{\delta_m}} \gamma(t_1) \dots \gamma(t_m) \\ &\quad \cdot \sum_{k_1} \alpha_{jk_1} \Phi_{k_1}(t_1) \dots \sum_{k_m} \alpha_{jk_m} \Phi_{k_m}(t_m) \\ &\quad \cdot \sum_{K_1, \dots, K_m} f_{K_1 \dots K_m}^J(y_0) F^{K_1}(t_1)(y_0) \dots F^{K_m}(t_m)(y_0) \\ &\quad + q \sum_{t_1 \in LT_{q-1}} \gamma(t_1) \sum_k \gamma_{jk} \Phi_k(t_1) \sum_K f_K^J(y_0) F^K(t_1)(y_0). \end{aligned}$$

The one-to-one correspondence between the summation set $\{(u, t_1, \dots, t_m) | u \in LS_q, t_j \in LT_{\delta_j}\}$ and LT_q together with the recursion formulas (7.9), (II.2.17), (II.2.18) now yields the result. \square

Comparing Theorems 7.3 and II.2.6 we obtain:

Table 7.1. Trees and order conditions up to order 5

$\varrho(t)$	t	graph	$\gamma(t)$	$\Phi_j(t)$	$p_t(\gamma)$
1	τ		1	1	1
2	t_{21}		2	$\sum_k \beta_{jk}$	$1/2 - \gamma$
3	t_{31}		3	$\sum_{k,l} \alpha_{jk} \alpha_{jl}$	$1/3$
	t_{32}		6	$\sum_{k,l} \beta_{jk} \beta_{kl}$	$1/6 - \gamma + \gamma^2$
4	t_{41}		4	$\sum_{k,l,m} \alpha_{jk} \alpha_{jl} \alpha_{jm}$	$1/4$
	t_{42}		8	$\sum_{k,l,m} \alpha_{jk} \beta_{kl} \alpha_{jm}$	$1/8 - \gamma/3$
	t_{43}		12	$\sum_{k,l,m} \beta_{jk} \alpha_{kl} \alpha_{km}$	$1/12 - \gamma/3$
	t_{44}		24	$\sum_{k,l,m} \beta_{jk} \beta_{kl} \beta_{lm}$	$1/24 - \gamma/2 + 3\gamma^2/2 - \gamma^3$
5	t_{51}		5	$\sum \alpha_{jk} \alpha_{jl} \alpha_{jm} \alpha_{jp}$	$1/5$
	t_{52}		10	$\sum \alpha_{jk} \beta_{kl} \alpha_{jm} \alpha_{jp}$	$1/10 - \gamma/4$
	t_{53}		15	$\sum \alpha_{jk} \alpha_{kl} \alpha_{km} \alpha_{jp}$	$1/15$
	t_{54}		30	$\sum \alpha_{jk} \beta_{kl} \beta_{lm} \alpha_{jp}$	$1/30 - \gamma/4 + \gamma^2/3$
	t_{55}		20	$\sum \alpha_{jk} \beta_{kl} \alpha_{jm} \beta_{mp}$	$1/20 - \gamma/4 + \gamma^2/3$
	t_{56}		20	$\sum \beta_{jk} \alpha_{kl} \alpha_{km} \alpha_{kp}$	$1/20 - \gamma/4$
	t_{57}		40	$\sum \beta_{jk} \alpha_{kl} \beta_{lm} \alpha_{kp}$	$1/40 - 5\gamma/24 + \gamma^2/3$
	t_{58}		60	$\sum \beta_{jk} \beta_{kl} \alpha_{lm} \alpha_{lp}$	$1/60 - \gamma/6 + \gamma^2/3$
	t_{59}		120	$\sum \beta_{jk} \beta_{kl} \beta_{lm} \beta_{mp}$	$1/120 - \gamma/6 + \gamma^2 - 2\gamma^3 + \gamma^4$

Theorem 7.4. A Rosenbrock method (7.4) with $J = f'(y_0)$ is of order p iff

$$\sum_j b_j \Phi_j(t) = \frac{1}{\gamma(t)} \quad \text{for } \varrho(t) \leq p. \quad (7.11)$$

□

The expressions $\Phi_j(t)$ simplify, if we introduce the abbreviation

$$\beta_{ij} = \alpha_{ij} + \gamma_{ij}. \quad (7.12)$$

The order conditions (7.11) for all trees up to order 5 are given in Table 7.1.

A further simplification of the order conditions (7.11) is possible if

$$\gamma_{ii} = \gamma \quad \text{for all } i \quad (7.13)$$

(It is unfortunate that in the current literature the letter γ is used for the parameter in (7.4) as well as for $\gamma(t)$ in (7.11) and we hope that no confusion will arise). In the same way as for DIRK methods, the summations in the expressions for $\Phi_j(t)$

in the 5th column of Table 7.1 again contain *more* terms than the corresponding expressions for explicit Runge-Kutta methods, since the matrix γ_{ij} (and hence β_{ij}) contains non-zero elements in the diagonal. The difference is that here these diagonal γ appear only for singly-branched vertices (see Definition 7.2). Therefore the procedure explained in Sect. IV.6 (see Formulas (6.1') and (6.1'')) must be slightly modified and leads to order conditions of the form

$$\sum_j {}'b_j \Phi_j(t) = p_t(\gamma) \quad (7.11')$$

where the polynomials $p_t(\gamma)$ are listed in the last column of Table 7.1.

The Stability Function

If we apply Method (7.4) to the test equation $y' = \lambda y$ and if we assume $J = f'(y_0) = \lambda$ then the numerical solution becomes $y_1 = R(h\lambda)y_0$ with

$$R(z) = 1 + zb^T(I - zB)^{-1} \mathbb{1} \quad (7.14)$$

where we have used the notation $b^T = (b_1, \dots, b_s)$ and $B = (\beta_{ij})_{i,j=1}^s$. Since B is a lower triangular matrix, the stability function (7.14) is equal to that of a DIRK-method with RK-matrix B . Properties of such stability functions have already been investigated in Sect. IV.6.

Construction of Methods of Order 4

In order to construct 4-stage Rosenbrock methods of order 4 we list, for convenience, the whole set of order conditions (c.f. Table 7.1.).

$$\bullet \quad b_1 + b_2 + b_3 + b_4 = 1 \quad (7.15a)$$

$$/ \quad b_2\beta'_2 + b_3\beta'_3 + b_4\beta'_4 = \frac{1}{2} - \gamma = p_{21}(\gamma) \quad (7.15b)$$

$$\vee \quad b_2\alpha_2^2 + b_3\alpha_3^2 + b_4\alpha_4^2 = \frac{1}{3} \quad (7.15c)$$

$$\rangle \quad b_3\beta_{32}\beta'_2 + b_4(\beta_{42}\beta'_2 + \beta_{43}\beta'_3) = \frac{1}{6} - \gamma + \gamma^2 = p_{32}(\gamma) \quad (7.15d)$$

$$\swarrow \quad b_2\alpha_2^3 + b_3\alpha_3^3 + b_4\alpha_4^3 = \frac{1}{4} \quad (7.15e)$$

$$\searrow \quad b_3\alpha_3\alpha_{32}\beta'_2 + b_4\alpha_4(\alpha_{42}\beta'_2 + \alpha_{43}\beta'_3) = \frac{1}{8} - \frac{\gamma}{3} = p_{42}(\gamma) \quad (7.15f)$$

$$\times \quad b_3\beta_{32}\alpha_2^2 + b_4(\beta_{42}\alpha_2^2 + \beta_{43}\alpha_3^2) = \frac{1}{12} - \frac{\gamma}{3} = p_{43}(\gamma) \quad (7.15g)$$

$$\langle \quad b_4\beta_{43}\beta_{32}\beta'_2 = \frac{1}{24} - \frac{\gamma}{2} + \frac{3}{2}\gamma^2 - \gamma^3 = p_{44}(\gamma) \quad (7.15h)$$

Here we have used the abbreviations

$$\alpha_i = \sum_{j=1}^{i-1} \alpha_{ij}, \quad \beta'_i = \sum_{j=1}^{i-1} \beta_{ij}. \quad (7.16)$$

For the sake of step size control we also look for an embedded formula (Wolfbrandt 1977, Kaps & Rentrop 1979)

$$\hat{y}_1 = y_0 + \sum_{j=1}^s \hat{b}_j k_j \quad (7.17)$$

which uses the same k_j -values as (7.4), but has different weights. This method should have order 3, i.e., the four conditions (7.15a)–(7.15d) should be satisfied also for the \hat{b}_i . These equations constitute the linear system

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & \beta'_2 & \beta'_3 & \beta'_4 \\ 0 & \alpha_2^2 & \alpha_3^2 & \alpha_4^2 \\ 0 & 0 & \beta_{32}\beta'_2 & \sum' \beta_{4j}\beta'_j \end{pmatrix} \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \\ \hat{b}_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 - \gamma \\ 1/3 \\ 1/6 - \gamma + \gamma^2 \end{pmatrix}. \quad (7.18)$$

Whenever the matrix in (7.18) is regular, uniqueness of the solutions of the linear system implies $\hat{b}_i = b_i$ ($i = 1, \dots, 4$) and the approximation \hat{y}_1 cannot be used for step size control. We therefore have to require that the matrix (7.18) be singular, i.e.,

$$(\beta'_2 \alpha_4^2 - \beta'_4 \alpha_2^2) \beta_{32} \beta'_2 = (\beta'_2 \alpha_3^2 - \beta'_3 \alpha_2^2) \sum_{j=2}^3 \beta_{4j} \beta'_j. \quad (7.19)$$

This condition guarantees the existence of a 3rd order embedded method (7.17), whenever (7.15) possesses a solution. The computation of the coefficients α_{ij} , β_{ij} , γ , b_i satisfying (7.15), (7.16) and (7.19) is now done in the following steps:

Step 1. Choose $\gamma > 0$ such that the stability function (7.14) has desirable stability properties (c.f. Table 6.3).

Step 2. Choose $\alpha_2, \alpha_3, \alpha_4$ and b_1, b_2, b_3, b_4 in such a way that the three conditions (7.15a), (7.15c), (7.15e) are fulfilled. One obviously has four degrees of freedom in this choice. Observe that the (b_i, α_i) need not be the coefficients of a standard quadrature formula, since $\sum b_i \alpha_i = 1/2$ need not be satisfied.

Step 3. Take β_{43} as a free parameter and compute $\beta_{32}\beta'_2$ from (7.15h), then $(\beta_{42}\beta'_2 + \beta_{43}\beta'_3)$ from (7.15d). These expressions, inserted into (7.19) yield a second relation between $\beta'_2, \beta'_3, \beta'_4$ (the first one is (7.15b)). Eliminating $(b_4\beta_{42} + b_3\beta_{32})$ from (7.15d) and (7.15g) gives

$$b_4 \beta_{43} (\beta'_2 \alpha_3^2 - \beta'_3 \alpha_2^2) = \beta'_2 p_{43}(\gamma) - \alpha_2^2 p_{32}(\gamma),$$

a third linear relation for $\beta'_2, \beta'_3, \beta'_4$. The resulting linear system is regular iff $b_4 \beta_{43} \alpha_2 \gamma (3\gamma - 1) \neq 0$.

Step 4. Once the β'_i are known we can find β_{32} and β_{42} from the values of $\beta_{32}\beta'_2$, $(\beta_{42}\beta'_2 + \beta_{43}\beta'_3)$ obtained in Step 3.

Step 5. Choose $\alpha_{32}, \alpha_{42}, \alpha_{43}$ according to (7.15f). One has two degrees of freedom to do this. Finally, the values α_i, β'_i yield α_{i1}, β_{i1} via condition (7.16).

Table 7.2 Rosenbrock methods of order 4

method	γ	parameter choices	$A(\alpha)$ -stable	$ R(\infty) $
GRK4A (Kaps-Rentrop 79)	0.395	$\alpha_2 = 0.438, \alpha_3 = 0.87$ $b_4 = 0.25$	$\pi/2$	0.995
GRK4T (Kaps-Rentrop 79)	0.231	$\alpha_2 = 2\gamma, (7.22), b_3 = 0$	89.3°	0.454
Shampine (1982)	0.5	$\alpha_2 = 2\gamma, (7.22), b_3 = 0$	$\pi/2$	1/3
Veldhuizen (1984)	0.225708	$\alpha_2 = 2\gamma, (7.22), b_3 = 0$	89.5°	0.24
Veldhuizen (1984)	0.5	$\alpha_2 = 2\gamma, \alpha_3 = 0.5, b_3 = 0$	$\pi/2$	1/3
<i>L</i> -stable method	0.572816	$\alpha_2 = 2\gamma, (7.22), b_3 = 0$	$\pi/2$	0

Most of the popular Rosenbrock methods are special cases of this construction (see Table 7.2). Usually the remaining free parameters are chosen as follows: if we require

$$\alpha_{43} = 0, \quad \alpha_{42} = \alpha_{32} \quad \text{and} \quad \alpha_{41} = \alpha_{31} \quad (7.20)$$

then the argument of f in (7.4) is the same for $i = 3$ and $i = 4$. Hence, the number of function evaluations is reduced by one. Further free parameters can be determined so that several order conditions of order five are satisfied. Multiplying the condition (7.15g) with α_2 and subtracting it from the order condition for the tree t_{56} yields

$$b_4\beta_{43}\alpha_3^2(\alpha_3 - \alpha_2) = p_{56}(\gamma) - \alpha_2 p_{43}(\gamma). \quad (7.21)$$

This determines β_{43} . The order condition for t_{51} can also easily be fulfilled in Step 2. If $\alpha_3 = \alpha_4$ (see (7.20)) this leads to the restriction

$$\alpha_3 = \frac{1/5 - \alpha_2/4}{1/4 - \alpha_2/3}. \quad (7.22)$$

In Table 7.2 we collect some well-known methods. All of them satisfy (7.20) and (7.21) (Only exception: the second method of van Veldhuizen for $\gamma = 0.5$ has $\beta_{43} = 0$ instead of (7.21)). The definition of the remaining free parameters is given in the first two columns. The last columns indicate some properties of the stability function.

Higher Order Methods

As for explicit Runge-Kutta methods the construction of higher order methods is facilitated by the use of *simplifying assumptions*. First, the condition

$$\sum_{i=j}^s b_i \beta_{ij} = b_j (1 - \alpha_j), \quad j = 1, \dots, s \quad (7.23)$$

plays a role similar to that of (II.1.12) for explicit Runge-Kutta methods. It implies that the order condition of the left-hand tree in Fig. 7.1 is a consequence of the two on the right-hand side. A difference to Runge-Kutta methods is that here the vertex directly above the root has to be multiply-branched.

The second type of simplifying assumption is (with $\beta_k = \sum_{l=1}^k \beta_{kl}$)

$$\sum_{k=1}^{j-1} \alpha_{jk} \beta_k = \frac{\alpha_j^2}{2}, \quad j = 2, \dots, s. \quad (7.24)$$

It has an effect similar to that of (II.5.7). As a consequence of (7.24) the order conditions of the two trees in Fig. 7.2 are equivalent. Again the vertex marked by an arrow has to be multiply-branched.

The use of the above simplifying assumptions has been exploited by Kaps & Wanner (1981) for their construction of methods up to order 6. Still higher order methods would need generalizations of the above simplifying assumptions (in analogy to $C(\eta)$ and $D(\zeta)$ of Sect. II.7).



Fig. 7.1. Reduction with (7.23)

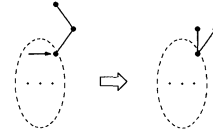


Fig. 7.2. Reduction with (7.24)

Implementation of Rosenbrock-Type Methods

A direct implementation of (7.4) requires, at each stage, the solution of a linear system with the matrix $I - h\gamma_{ii}J$ and also the matrix-vector multiplication $J \cdot \sum \gamma_{ij}k_j$. The latter can be avoided by the introduction of the new variables

$$u_i = \sum_{j=1}^i \gamma_{ij}k_j, \quad i = 1, \dots, s.$$

If $\gamma_{ii} \neq 0$ for all i , the matrix $\Gamma = (\gamma_{ij})$ is invertible and the k_i can be recovered from the u_i :

$$k_i = \frac{1}{\gamma_{ii}} u_i - \sum_{j=1}^{i-1} c_{ij} u_j, \quad C = \text{diag}(\gamma_{11}^{-1}, \dots, \gamma_{ss}^{-1}) - \Gamma^{-1}.$$

Inserting this formula into (7.4) and dividing by h yields

$$\begin{aligned} \left(\frac{1}{h\gamma_{ii}}I - J\right)u_i &= f\left(y_0 + \sum_{j=1}^{i-1} a_{ij}u_j\right) + \sum_{j=1}^{i-1} \left(\frac{c_{ij}}{h}\right)u_j, \quad i = 1, \dots, s \\ y_1 &= y_0 + \sum_{j=1}^s m_j u_j, \end{aligned} \quad (7.25)$$

where

$$(a_{ij}) = (\alpha_{ij})\Gamma^{-1}, \quad (m_1, \dots, m_s) = (b_1, \dots, b_s)\Gamma^{-1}.$$

Compared to (7.4) the formulation (7.25) of a Rosenbrock method avoids not only the above mentioned matrix-vector multiplication, but also the n^2 multiplications for $(\gamma_{ii}h)J$. Similar transformations were first proposed by Wolfbrandt (1977), Kaps & Wanner (1981) and Shampine (1982). The formulation (7.25) can be found in Kaps, Poon & Bui (1985).

For *non-autonomous* problems this transformation yields

$$\begin{aligned} \left(\frac{1}{h\gamma_{ii}}I - \frac{\partial f}{\partial y}(x_0, y_0)\right)u_i &= f\left(x_0 + \alpha_i h, y_0 + \sum_{j=1}^{i-1} a_{ij}u_j\right) \\ &+ \sum_{j=1}^{i-1} \left(\frac{c_{ij}}{h}\right)u_j + \gamma_i h \frac{\partial f}{\partial x}(x_0, y_0) \end{aligned} \quad (7.26)$$

with α_i and γ_i given by (7.5).

For *implicit differential equations* of the form (7.2b) the transformed Rosenbrock method becomes

$$\begin{aligned} \left(\frac{1}{h\gamma_{ii}}M - \frac{\partial f}{\partial y}(x_0, y_0)\right)u_i &= f\left(x_0 + \alpha_i h, y_0 + \sum_{j=1}^{i-1} a_{ij}u_j\right) \\ &+ M \sum_{j=1}^{i-1} \left(\frac{c_{ij}}{h}\right)u_j + \gamma_i h \frac{\partial f}{\partial x}(x_0, y_0). \end{aligned} \quad (7.27)$$

Coding. Rosenbrock methods are nearly as simple to implement as explicit Runge-Kutta methods. The only difference is that at each step the Jacobian $\partial f/\partial y$ has to be evaluated and s linear systems have to be solved. Thus, one can take an explicit RK code (say DOPRI5), add four lines which compute $\partial f/\partial y$ by finite differences (or call a user-supplied subroutine JAC which furnishes it analytically); add further a call to a Gaussian DEComposition routine, and add to each evaluation-stage a call to a linear SOLver. Since the method is of order 4(3), the step size prediction formula

$$h_{new} = h \cdot \min\left\{6., \max\left(0.2, 0.9 \cdot (Tol/err)^{1/4}\right)\right\} \quad (7.28)$$

seems appropriate.

However, we want the code to work economically for non-autonomous problems as well as for implicit equations. Further, if the dimension of the system is large, it becomes crucial that the linear algebra be done, whenever possible, in banded form. All these possibilities, autonomous or not, implicit or explicit, $\partial f/\partial y$ banded or not, B banded or not, $\partial f/\partial y$ analytic or not, (“... that is the question”) lead to 2^5 different cases, for each of which the code contains special parts for high efficiency. Needless to say, it works well on all stiff problems of Sect. IV.1. A more thorough comparison and testing will be given in Sect. IV.10.

The “Hump”

On some very stiff equations, however, the code shows a curious behaviour: consider the van der Pol equation in singular perturbation form (1.5') with

$$\varepsilon = 10^{-6}, \quad y_1(0) = 2, \quad y_2(0) = -0.66. \quad (7.29)$$

We further select method GRK4T (Table 7.2; each other method there behaves similarly) and $Tol = 7 \cdot 10^{-5}$. Fig. 7.3 shows the numerical solution y_1 as well as the step sizes chosen by the code. There all rejected steps are indicated by an \times .

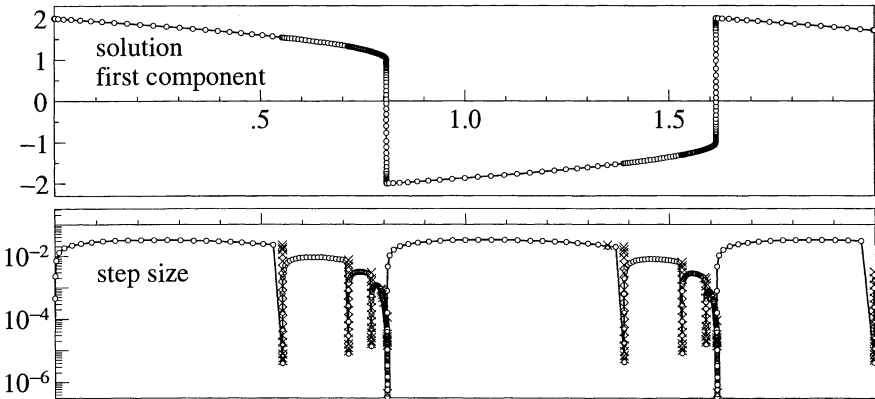


Fig. 7.3. Step sizes for GRK4T at Equation (1.5')

Curious step size drops (by a factor of about 10^{-3}) occur without any apparent exterior reason. Further, these drops are accompanied by a huge number of step rejections (up to 20). In order to understand this phenomenon, we present in the left picture of Fig. 7.4 the *exact local error* as well as the *estimated local error* $\|y_1 - \hat{y}_1\|$ at $x = 0.55139$ as a function of the step size h (both in logarithmic scale). The current step size is marked by large symbols. The error behaves like $C \cdot h^5$ only for very small h ($\leq 10^{-6} = \varepsilon$). Between $h = 10^{-5}$ and the step size actually used ($\approx 10^{-2}$) the error is more or less constant. Whenever this constant is larger than Tol (horizontal broken line), the code is forced to decrease the step

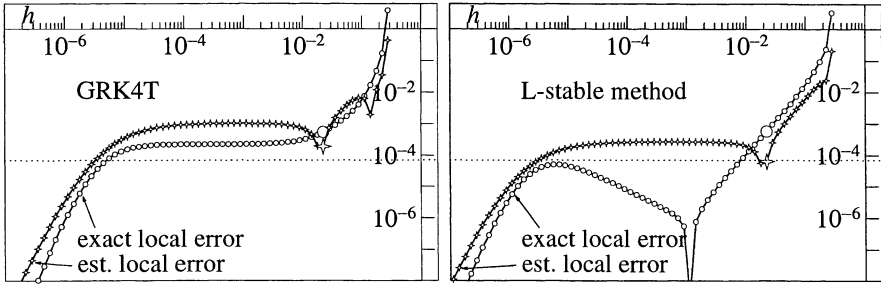


Fig. 7.4. Study of local error for (1.5') at $x = 0.55139$

size until $h \approx \varepsilon$. As a first remedy, we accelerate this lengthy process, as Shampine (1982) also did, by more drastical step size reductions ($h_{new} = h/10$) after each second consecutive step rejection. It also turns out (see right picture of Fig. 7.4) that the effect disappears in the neighbourhood of the actual step size for the L -stable method (where $R(\infty) = 0$). Methods with $R(\infty) = 0$ and also $\hat{R}(\infty) = 0$ have been derived by Kaps & Ostermann (1990).

A more thorough understanding of these phenomena is possible by the consideration of singular perturbation problems (Chapter VI).

Methods with Inexact Jacobian (W -Methods)

The relevant question is now, what is the cheapest type of implicitness we have to require. (Steihaug & Wolfbrandt 1979)

All the above theory is built on the assumption that J is the exact Jacobian $\partial f / \partial y$. This implies that the matrix must be evaluated at every step, which can make the computations costly. The following attempt, due to Steihaug & Wolfbrandt (1979), searches for order conditions which assure classical order for *all* approximations A of $\partial f / \partial y$. The latter is then maintained over several steps and is just used to assure stability. The derivation of the order conditions must now be done somewhat differently: if J is replaced by an arbitrary matrix A , Formula (7.6) becomes

$$(k_j^J)^{(q)}|_{h=0} = q(f^J(g_j))^{(q-1)}|_{h=0} + q \sum_K A_K^J \sum_k \gamma_{jk} (k_k^K)^{(q-1)}|_{h=0} \quad (7.30)$$

where $A = (A_K^J)_{J,K=1}^n$, and we obtain

$$(k_j^J)^{(2)}|_{h=0} = 2 \sum_K f_K^J f^K \sum_k \alpha_{jk} + 2 \sum_K A_K^J f^K \sum_k \gamma_{jk}. \quad (7.31;2)$$

Inserted into (7.8), the first term must equal the derivative of the exact solution and

the second must be zero. Similarly, we obtain instead of (7.7;3)

$$\begin{aligned}
 (k_j^J)^{(3)}|_{h=0} &= 3 \sum_{K,L} f_{KL}^J f^K f^L \sum_{k,l} \alpha_{jk} \alpha_{jl} \\
 &+ 3 \cdot 2 \sum_{K,L} f_K^J f_L^K f^L \sum_{k,l} \alpha_{jk} \alpha_{kl} + 3 \cdot 2 \sum_{K,L} f_K^J A_L^K f^L \sum_{k,l} \alpha_{jk} \gamma_{kl} \\
 &+ 3 \cdot 2 \sum_{K,L} A_K^J f_L^K f^L \sum_{k,l} \gamma_{jk} \alpha_{kl} + 3 \cdot 2 \sum_{K,L} A_K^J A_L^K f^L \sum_{k,l} \gamma_{jk} \gamma_{kl}
 \end{aligned} \tag{7.31;3}$$

and the order conditions for order three become

$$\begin{array}{ll}
 \bullet_j & \sum b_j = 1 \\
 \begin{array}{c} \nearrow^k \\ \bullet_j \end{array} & \sum b_j \alpha_{jk} = 1/2 \\
 \begin{array}{c} \nearrow^k \\ \circ_j \end{array} & \sum b_j \gamma_{jk} = 0 \\
 \begin{array}{c} \nearrow^k \searrow^l \\ \bullet_j \end{array} & \sum b_j \alpha_{jk} \alpha_{jl} = 1/3 \\
 \begin{array}{c} \nearrow^k \searrow^l \\ \circ_j \end{array} & \sum b_j \alpha_{jk} \alpha_{kl} = 1/6 \\
 \begin{array}{c} \nearrow^l \searrow^k \\ \bullet_j \end{array} & \sum b_j \alpha_{jk} \gamma_{kl} = 0 \\
 \begin{array}{c} \nearrow^l \searrow^k \\ \circ_j \end{array} & \sum b_j \gamma_{jk} \alpha_{kl} = 0 \\
 \begin{array}{c} \nearrow^l \searrow^k \\ \circ_j \end{array} & \sum b_j \gamma_{jk} \gamma_{kl} = 0.
 \end{array} \tag{7.32}$$

For a graphical representation of the elementary differentials in (7.31;q) and of the order conditions (7.32) we need trees with two different kinds of vertices (one representing f and the other A). As in Sect. II.15 we use “meagre” and “fat” vertices (see Definitions II.15.1 to II.15.4). Not all trees with meagre and fat vertices (P -trees) have to be considered. From the above derivation we see that fat vertices have to be singly-branched (derivatives of the constant matrix A are zero) and that they cannot be at the end of a branch. We therefore use the notation

$$TW = \{ P\text{-trees ; end-vertices are meagre and fat vertices are singly-branched} \} \tag{7.33}$$

and if the vertices are labelled monotonically, we write LTW .

Definition 7.5. The *elementary differentials* for trees $t \in TW$ are defined recursively by $F^J(\tau)(y) = f^J(y)$ and

$$F^J(t)(y) = \begin{cases} \sum_{K_1, \dots, K_m} f_{K_1, \dots, K_m}^J(y) \cdot \left(F^{K_1}(t_1)(y), \dots, F^{K_m}(t_m)(y) \right) & \text{if } t = {}_a[t_1, \dots, t_m] \quad (\text{meagre root}) \\ \sum_K A_K^J \cdot F^K(t_1)(y) & \text{if } t = {}_b[t_1] \quad (\text{fat root}). \end{cases}$$

Definition 7.6. For $t \in TW$ we let $\Phi_j(\tau) = 1$ and

$$\Phi_j(t) = \begin{cases} \sum_{k_1, \dots, k_m} \alpha_{jk_1} \dots \alpha_{jk_m} \Phi_{k_1}(t_1) \dots \Phi_{k_m}(t_m) & \text{if } t = {}_a[t_1, \dots, t_m] \\ \sum_k \gamma_{jk} \Phi_k(t_1) & \text{if } t = {}_b[t_1]. \end{cases}$$

We remark that T (the set of trees as considered for Runge-Kutta methods) is a subset of TW and that the above definitions coincide with Definitions II.2.3 and II.2.9 (c.f. also Formulas (II.2.18) and (II.2.19)). The general result is now the following

Theorem 7.7. A W -method (7.4) with $J = A$ arbitrary is of order p iff

$$\begin{aligned} \sum_j b_j \Phi_j(t) &= \frac{1}{\gamma(t)} & \text{for } t \in T \text{ with } \varrho(t) \leq p, \text{ and} \\ \sum_j b_j \Phi_j(t) &= 0 & \text{for } t \in TW \setminus T \text{ with } \varrho(t) \leq p. \end{aligned}$$

The *proof* is essentially the same as for Theorems 7.3 and 7.4. □

Table 7.3. Number of order conditions for W -methods

order p	1	2	3	4	5	6	7	8
no. of conditions	1	3	8	21	58	166	498	1540

The number of order conditions for W -methods is rather large (see Table 7.3), since each tree of T with κ singly-branched vertices gives rise to 2^κ order conditions (in the case of symmetry some may be identical). Therefore, W -methods of higher order are best obtained by extrapolation (see Sect. IV.9).

The *stability* investigation for linearly implicit methods with $A \neq f'(y_0)$ is very complicated. If we linearize the differential equation (as in the beginning of Sect. IV.2) and assume the Jacobian to be constant, we arrive at a recursion of the form

$$y_1 = R(hf'(y_0), hA)y_0.$$

Since, in general, the matrices $f'(y_0)$ and A cannot be diagonalized simultaneously, the consideration of scalar test equations is not justified. Stability investigations for the case when $\|f'(y_0) - A\|$ is small will be considered in Sect. IV.11.

Exercises

1. (Kaps 1977). There exists no Rosenbrock method (7.4) with $s = 4$ and $p = 5$. Prove this.
2. (Nørsett & Wolfbrandt 1979). Generalize the derivation of order conditions for Runge-Kutta methods with the help of B-series (Sect. II.11, page 247) to Rosenbrock methods.

Hint. Prove that, for a B-series $B(\mathbf{a}, y_0)$ with $\mathbf{a}: T \rightarrow \mathbb{R}$ satisfying $\mathbf{a}(\emptyset) = 0$,

$$hf'(y_0)B(\mathbf{a}, y_0) = B(\hat{\mathbf{a}}, y_0)$$

is again a B-series with coefficients

$$\hat{\mathbf{a}}(t) = \begin{cases} \varrho(t)\mathbf{a}(t_1) & \text{if } t = [t_1] \\ 0 & \text{else.} \end{cases}$$

3. Cooper & Sayfy (1983) consider *additive* Runge-Kutta methods

$$\begin{aligned} g_i &= y_0 + h \sum_{j=1}^{i-1} \alpha_{ij} f(x_0 + c_j h, g_j) + hJ \sum_{j=1}^i \eta_{ij} g_j \quad i = 1, \dots, s+1 \\ y_1 &= g_{s+1} \end{aligned} \tag{7.34}$$

whose coefficients satisfy $\sum_{j=1}^{i-1} \alpha_{ij} = c_i$, $\sum_{j=1}^i \eta_{ij} = 0$.

a) Prove that (7.34) is equivalent to (7.4) whenever $\alpha_{s+1,i} = b_i$ and

$$(\eta_{ij})(\alpha_{ij}) = (\alpha_{ij})(\gamma_{ij}). \tag{7.35}$$

Here all matrices are of dimension $(s+1) \times (s+1)$. The last line of (γ_{ij}) need not be specified since the last column of (α_{ij}) is zero.

b) If the coefficients of (7.34) satisfy $\alpha_{i,i-1} \neq 0$ for all i , then we can always find an equivalent method of type (7.4).

4. (Verwer 1980, Verwer & Scholz 1983). Derive order conditions for Rosenbrock methods “with time-lagged Jacobian”, i.e., methods of type (7.4) where J is assumed to be $f'(y(x_0 - \omega h))$. If ω is the step ratio h_{old}/h , this allows re-use of the Jacobian of the previous step.
5. (Kaps & Ostermann 1989). Show that some order conditions of (7.32) can be shifted to higher orders if it is assumed that

$$f'(y_0) - J = \mathcal{O}(h).$$

This makes the conditions of Exercise 4 independent of ω .

Result. The number of order-shifts is equal to the number of fat nodes.

IV.8 Implementation of Implicit Runge-Kutta Methods

These have not been used to any great extent . . .
(S.P. Nørsett 1976)

However, the implementation difficulties of these methods have
precluded their general use; . . . (J.M. Varah 1979)

Although Runge-Kutta methods present an attractive alternative,
especially for stiff problems, . . . it is generally believed that they
will never be competitive with multistep methods.
(K. Burrage, J.C. Butcher & F.H. Chipman 1980)

Runge-Kutta methods for stiff problems, we are just beginning to
explore them . . . (L. Shampine in Aiken 1985)

If the dimension of the differential equation $y' = f(x, y)$ is n , then the s -stage fully implicit Runge-Kutta method (3.1) involves a $n \cdot s$ -dimensional nonlinear system for the unknowns g_1, \dots, g_s . An efficient solution of this system is the main problem in the implementation of an implicit Runge-Kutta method.

Among the methods discussed in Sect. IV.5, the processes Radau IIA of Ehle, which are L -stable and of high order, seem to be particularly promising. Most of the questions arising (starting values and stopping criteria for the simplified Newton iterations, efficient solution of the linear systems, and the selection of the step sizes) are discussed here for the particular Ehle method with $s = 3$ and $p = 5$. This then constitutes a description of the code RADAU5 of the appendix. An adaptation of the described techniques to other fully implicit Runge-Kutta methods is more or less straight-forward, if the Runge-Kutta matrix has at least one real eigenvalue. We also describe briefly our implementation of the diagonal implicit method SDIRK4 (Formula (6.16)).

Reformulation of the Nonlinear System

In order to reduce the influence of round-off errors we prefer to work with the smaller quantities

$$z_i = g_i - y_0. \quad (8.1)$$

Then (3.1a) becomes

$$z_i = h \sum_{j=1}^s a_{ij} f(x_0 + c_j h, y_0 + z_j) \quad i = 1, \dots, s. \quad (8.2a)$$

Whenever the solution z_1, \dots, z_s of the system (8.2a) is known, then (3.1b) is an explicit formula for y_1 . A direct application of this requires s additional function evaluations. These can be avoided, if the matrix $A = (a_{ij})$ of the Runge-Kutta

coefficients is nonsingular. Indeed, (8.2a) can be written as

$$\begin{pmatrix} z_1 \\ \vdots \\ z_s \end{pmatrix} = A \begin{pmatrix} hf(x_0 + c_1 h, y_0 + z_1) \\ \vdots \\ hf(x_0 + c_s h, y_0 + z_s) \end{pmatrix},$$

so that (3.1b) is seen to be equivalent to

$$y_1 = y_0 + \sum_{i=1}^s d_i z_i \quad (8.2b)$$

where

$$(d_1, \dots, d_s) = (b_1, \dots, b_s)A^{-1}. \quad (8.3)$$

For the 3-stage Radau IIA method (Table 5.6) the vector d is simply $(0, 0, 1)$, since $b_i = a_{si}$ for all i .

Another advantage of Formula (8.2b) is the following: the quantities z_1, \dots, z_s are computed iteratively and are therefore affected by iteration errors. The evaluation of $f(x_0 + c_i h, y_0 + z_i)$ in Eq. (3.1b) would then, due to the large Lipschitz constant of f , amplify these errors, which then “can be disastrously inaccurate for a stiff problem” (L.F. Shampine 1980).

Simplified Newton Iterations

For a general nonlinear differential equation the system (8.2a) has to be solved iteratively. In the stone-age of stiff computation (i.e., before 1967) people were usually thinking of simple fixed-point iteration. But this transforms the algorithm into an explicit method and destroys the good stability properties. The paper of Liniger & Willoughby (1970) then showed the advantages of using Newton’s method for this purpose. Newton’s method applied to system (8.2a) needs for each iteration the solution of a linear system with matrix

$$\begin{pmatrix} I - ha_{11} \frac{\partial f}{\partial y}(x_0 + c_1 h, y_0 + z_1) & \dots & -ha_{1s} \frac{\partial f}{\partial y}(x_0 + c_s h, y_0 + z_s) \\ \vdots & & \vdots \\ -ha_{s1} \frac{\partial f}{\partial y}(x_0 + c_1 h, y_0 + z_1) & \dots & I - ha_{ss} \frac{\partial f}{\partial y}(x_0 + c_s h, y_0 + z_s) \end{pmatrix}.$$

In order to simplify this, we replace all Jacobians $\frac{\partial f}{\partial y}(x_0 + c_i h, y_0 + z_i)$ by an approximation

$$J \approx \frac{\partial f}{\partial y}(x_0, y_0).$$

Then the simplified Newton iterations for (8.2a) become

$$\begin{aligned} (I - hA \otimes J) \Delta Z^k &= -Z^k + h(A \otimes I)F(Z^k) \\ Z^{k+1} &= Z^k + \Delta Z^k. \end{aligned} \quad (8.4)$$

Here $Z^k = (z_1^k, \dots, z_s^k)^T$ is the k -th approximation to the solution, and $\Delta Z^k = (\Delta z_1^k, \dots, \Delta z_s^k)^T$ are the increments. $F(Z^k)$ is an abbreviation for

$$F(Z^k) = (f(x_0 + c_1 h, y_0 + z_1^k), \dots, f(x_0 + c_s h, y_0 + z_s^k))^T.$$

Each iteration requires s evaluations of f and the solution of a $n \cdot s$ -dimensional linear system. The matrix $(I - hA \otimes J)$ is the same for all iterations. Its LU-decomposition is done only once and is usually very costly.

Starting Values for the Newton Iteration. A natural and simple choice for the starting values in the iteration (8.4) (or equivalently (8.13) below), since the exact solution of (8.2a) satisfies $z_i = \mathcal{O}(h)$, would be

$$z_i^0 = 0, \quad i = 1, \dots, s. \quad (8.5)$$

However, better choices are possible in general. If the implicit Runge-Kutta method satisfies the condition $C(\eta)$ (see Sections IV.5 and II.7) for some $\eta \leq s$, then

$$z_i = y(x_0 + c_i h) - y_0 + \mathcal{O}(h^{\eta+1}). \quad (8.6)$$

Suppose now that $c_i \neq 0$ ($i = 1, \dots, s$) and consider the interpolation polynomial of degree s , defined by

$$q(0) = 0, \quad q(c_i) = z_i \quad i = 1, \dots, s.$$

Since the interpolation error is of size $\mathcal{O}(h^{s+1})$ we obtain together with (8.6)

$$y(x_0 + th) - y_0 - q(t) = \mathcal{O}(h^{\eta+1})$$

(cf. Theorem 7.10 of Chapter II for collocation methods). We use the values of $q(t)$ also beyond the interval $[0, 1]$ and take

$$z_i^0 = q(1 + w c_i) + y_0 - y_1, \quad i = 1, \dots, s, \quad w = h_{\text{new}}/h_{\text{old}} \quad (8.5')$$

as starting values for the Newton iteration in the subsequent step. Numerical experiments with the 3-stage Radau IIA method have shown that (8.5') usually leads to a faster convergence than (8.5).

Stopping Criterion. This question is closely related to an estimation of the iteration error. Since convergence is linear, we have

$$\|\Delta Z^{k+1}\| \leq \Theta \|\Delta Z^k\|, \quad \text{hopefully with } \Theta < 1. \quad (8.7)$$

Applying the triangle inequality to

$$Z^{k+1} - Z^* = (Z^{k+1} - Z^{k+2}) + (Z^{k+2} - Z^{k+3}) + \dots$$

(where Z^* is the exact solution of (8.2a)) yields the estimate

$$\|Z^{k+1} - Z^*\| \leq \frac{\Theta}{1 - \Theta} \|\Delta Z^k\|. \quad (8.8)$$

The convergence rate Θ can be estimated by the computed quantities

$$\Theta_k = \|\Delta Z^k\| / \|\Delta Z^{k-1}\|, \quad k \geq 1. \quad (8.9)$$

It is clear that the iteration error should not be larger than the local discretization error, which is usually kept close to Tol . We therefore stop the iteration when

$$\eta_k \|\Delta Z^k\| \leq \kappa \cdot Tol \quad \text{with} \quad \eta_k = \frac{\Theta_k}{1 - \Theta_k} \quad (8.10)$$

and accept Z^{k+1} as approximation to Z^* . This strategy can only be applied after at least two iterations. In order to be able to stop the computations after the first iteration already (which is especially advantageous for linear systems) we take for $k = 0$ the quantity

$$\eta_0 = (\max(\eta_{old}, Uround))^{0.8}$$

where η_{old} is the last η_k of the preceding step. It remains to make a good choice for the parameter κ in (8.10). To this end we applied the code RADAU5 for many different values of κ between 10 and 10^{-4} and with some different tolerances Tol to several differential equations. The observation was that the code works most efficiently for values of κ around 10^{-1} or 10^{-2} .

It is our experience that the code becomes more efficient when we allow a relatively high number of iterations (e.g., $k_{max} = 7$ or 10). During these k_{max} iterations, the computations are interrupted and restarted with a smaller stepsize (for example with $h := h/2$) if one of the following situations occurs

- a) there is a k with $\Theta_k \geq 1$ (the iteration “diverges”);
- b) for some k ,

$$\frac{\Theta_k^{k_{max}-k}}{1 - \Theta_k} \|\Delta Z^k\| > \kappa \cdot Tol. \quad (8.11)$$

The left-hand expression in (8.11) is a rough estimate of the iteration error to be expected after $k_{max} - 1$ iterations. The norm, used in all these formulas, should be the same as the one used for the local error estimator.

If only one Newton iteration was necessary to satisfy (8.10) or if the last Θ_k was very small, say $\leq 10^{-3}$, then we don't recompute the Jacobian in the next step. As a consequence, the Jacobian is computed only once for linear problems with constant coefficients (as long as no step rejection occurs).

The Linear System

An essential gain of numerical work for the solution of the linear system (8.4) is obtained by the following method, introduced independently by Butcher (1976) and Bickart (1977), which exploits with much profit the special structure of the matrix $I - hA \otimes J$ in (8.4).

The idea is to premultiply (8.4) by $(hA)^{-1} \otimes I$ (we suppose here that A is invertible) and to transform A^{-1} to a simple matrix (diagonal, block diagonal, triangular or Jordan canonical form)

$$T^{-1} A^{-1} T = \Lambda. \quad (8.12)$$

With the transformed variables $W^k = (T^{-1} \otimes I)Z^k$, the iteration (8.4) becomes equivalent to

$$\begin{aligned} (h^{-1}\Lambda \otimes I - I \otimes J)\Delta W^k &= -h^{-1}(\Lambda \otimes I)W^k + (T^{-1} \otimes I)F((T \otimes I)W^k) \\ W^{k+1} &= W^k + \Delta W^k. \end{aligned} \quad (8.13)$$

We also replace Z^k and ΔZ^k by W^k and ΔW^k in the formulas (8.7)–(8.11) (and thereby again save some work).

For the sequel, we suppose that the matrix A^{-1} has one real eigenvalue $\hat{\gamma}$ and one complex conjugate eigenvalue pair $\hat{\alpha} \pm i\hat{\beta}$. This is a typical situation for 3-stage implicit Runge-Kutta methods such as Radau IIA. With $\gamma = h^{-1}\hat{\gamma}$, $\alpha = h^{-1}\hat{\alpha}$, $\beta = h^{-1}\hat{\beta}$ the matrix in (8.13) becomes

$$\begin{pmatrix} \gamma I - J & 0 & 0 \\ 0 & \alpha I - J & -\beta I \\ 0 & \beta I & \alpha I - J \end{pmatrix} \quad (8.14)$$

so that (8.13) splits into two linear systems of dimension n and $2n$, respectively. Several ideas are possible to exploit the special structure of the $2n \times 2n$ -submatrix. The easiest and numerically most stable way has turned out to be the following: transform the real subsystem of dimension $2n$ into an n -dimensional, complex system

$$((\alpha + i\beta)I - J)(u + iv) = a + ib \quad (8.14')$$

and apply simple Gaussian elimination. For machines without complex arithmetic, one just has to modify the linear algebra routines. Then a complex multiplication consists of 4 real multiplications and the amount of work for the solution of (8.14') becomes approximately $4n^3/3$ operations. Thus the total work for system (8.14) is about $5n^3/3$ operations. Compared to $(3n)^3/3$, which would be the number of operations necessary for decomposing the untransformed matrix $I - hA \otimes J$ in (8.4), we gain a factor of about 5 in arithmetical operations. Observe that the transformations, such as $Z^k = (T \otimes I)W^k$, need only $\mathcal{O}(n)$ additions and multiplications. The gain is still more drastic for methods with more than 3 stages.

Transformation to Hessenberg Form. For large systems with a full Jacobian J a further gain is possible by transforming J to Hessenberg form

$$S^{-1}JS = H = \begin{pmatrix} * & \dots & * & * \\ * & & & * \\ & \ddots & & \vdots \\ & & * & * \end{pmatrix}. \quad (8.15)$$

This procedure was originally proposed for multistep methods by Enright (1978) and extended to the Runge-Kutta case by Varah (1979). With the code ELMHES, taken from LINPACK, this is performed with $2n^3/3$ operations. Because the multiplication of S with a vector needs only $n^2/2$ operations (observe that S is triangular) the solution of (8.13) is found in $\mathcal{O}(n^2)$ operations, if the Hessenberg matrix H is known. This transformation is especially advantageous, if the Jacobian J is not changed during several steps.

Step Size Selection

One possibility to select the step sizes is Richardson extrapolation (cf. Sect. II.4). We describe here the use of an embedded pair of methods which is easier to program and which makes the code more flexible. The following formulas are for the special case of the 3-stage Radau IIA methods; the same ideas are applicable to all implicit Runge-Kutta methods, whose Runge-Kutta matrix has at least one real eigenvalue.

Embedded Formula. Since our method is of optimal order, it is impossible to embed it efficiently into one of still higher order. Therefore we search for a lower order method of the form

$$\hat{y}_1 = y_0 + h \left(\hat{b}_0 f(x_0, y_0) + \sum_{i=1}^3 \hat{b}_i f(x_0 + c_i h, g_i) \right) \quad (8.16)$$

where g_1, g_2, g_3 are the values obtained from the Radau IIA method and $\hat{b}_0 \neq 0$ (the choice $\hat{b}_0 = \gamma_0 = \hat{\gamma}^{-1}$, where $\hat{\gamma}$ is the real eigenvalue of the matrix A^{-1} , again saves some multiplications). The difference

$$\hat{y}_1 - y_1 = \gamma_0 h f(x_0, y_0) + \sum_{i=1}^3 (\hat{b}_i - b_i) h f(x_0 + c_i h, g_i),$$

which can also be written in the form

$$\hat{y}_1 - y_1 = \gamma_0 h f(x_0, y_0) + e_1 z_1 + e_2 z_2 + e_3 z_3, \quad (8.17)$$

then serves for error estimation. In order that $\hat{y}_1 - y_1 = \mathcal{O}(h^4)$ the coefficients have to satisfy

$$(e_1, e_2, e_3) = \frac{\gamma_0}{3} (-13 - 7\sqrt{6}, -13 + 7\sqrt{6}, -1). \quad (8.18)$$

Unfortunately, for $y' = \lambda y$ and $h\lambda \rightarrow \infty$ the difference (8.17) behaves like $\hat{y}_1 - y_1 \approx \gamma_0 h \lambda y_0$, which is unbounded and therefore not suitable for stiff equations. We propose (an idea of Shampine) to use instead

$$err = (I - h\gamma_0 J)^{-1} (\hat{y}_1 - y_1). \quad (8.19)$$

The LU-decomposition of $((h\gamma_0)^{-1}I - J)$ is available anyway from the previous work, so that the computation of (8.19) is cheap. For $h \rightarrow 0$ we still have $err = \mathcal{O}(h^4)$, and for $h\lambda \rightarrow \infty$ (if $y' = \lambda y$ and $J = \lambda$) we obtain $err \rightarrow -1$.

This behaviour (for $h\lambda \rightarrow \infty$) is already much better than that for $\hat{y}_1 - y_1$, but it is not good enough in order to avoid the “hump” phenomenon, described in Sect. IV.7. In the first step and after every rejected step for which $\|err\| > 1$, we therefore use instead of (8.19) the expression

$$\widetilde{err} = (I - h\gamma_0 J)^{-1} (\gamma_0 h f(x_0, y_0 + err) + e_1 z_1 + e_2 z_2 + e_3 z_3) \quad (8.20)$$

for step size prediction. This requires one additional function evaluation, but satisfies $\widetilde{err} \rightarrow 0$ for $h\lambda \rightarrow \infty$, as does the error of the numerical solution.

Standard Step Size Controller. Since the expressions (8.19) and (8.20) behave like $\mathcal{O}(h^4)$ for $h \rightarrow 0$, the standard step size prediction leads to

$$h_{\text{new}} = fac \cdot h_{\text{old}} \cdot \|err\|^{-1/4}. \quad (8.21)$$

where

$$\|err\| = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\frac{err_i}{sc_i} \right)^2},$$

and $sc_i = Atol_i + \max(|y_{0i}|, |y_{1i}|) \cdot Rtol_i$ as in (4.11) of Sect. II.4. Here, the safety factor fac is proposed to depend on $Newt$, the number of Newton iterations of the current step and on the maximal number of Newton iterations k_{max} , say, as: $fac = 0.9 \times (2k_{\text{max}} + 1) / (2k_{\text{max}} + Newt)$.

In order to save LU-decompositions of the matrix (8.14), we also include the following strategy: if no Jacobian is recomputed and if the step size h_{new} , defined by (8.21), satisfies

$$c_1 h_{\text{old}} \leq h_{\text{new}} \leq c_2 h_{\text{old}} \quad (8.22)$$

with, say $c_1 = 1.0$ and $c_2 = 1.2$, then we retain h_{old} for the following step.

Predictive Controller. The step size prediction by formula (8.21) has the disadvantage that step size reductions by more than the factor fac are not possible without step rejections (observe that $h_{\text{new}} < fac \cdot h_{\text{old}}$ implies $\|err\| > 1$). For stiff differential equations, however, a rapid decrease of the step size is often required (see for example the situation of Fig. 8.1, where the step size drops from 10^{-2} to 10^{-7} within a very small time interval). Denoting by err_{n+1} the error expression (8.19) (or (8.20)), computed in the n th step with step size h_n , step size predictions are typically derived from the asymptotic formula

$$\|err_{n+1}\| = C_n h_n^4. \quad (8.23)$$

The strategy (8.21) is based on the additional assumption $C_{n+1} \approx C_n$, which, as we have seen, is not always very realistic.

A careful control-theoretic study of step size strategies has been undertaken by Gustafsson (1994). He came to the conclusion that a better model is to assume that $\log C_n$ is a linear function of n . This means that $\log C_{n+1} - \log C_n$ is constant or, equivalently,

$$C_{n+1}/C_n \approx C_n/C_{n-1}. \quad (8.24)$$

Inserting C_n and C_{n-1} from (8.23) and C_{n+1} from $1 = C_{n+1} h_{\text{new}}^4$ into (8.24) yields

$$h_{\text{new}} = fac \cdot h_n \left(\frac{1}{\|err_{n+1}\|} \right)^{1/4} \cdot \frac{h_n}{h_{n-1}} \left(\frac{\|err_n\|}{\|err_{n+1}\|} \right)^{1/4}. \quad (8.25)$$

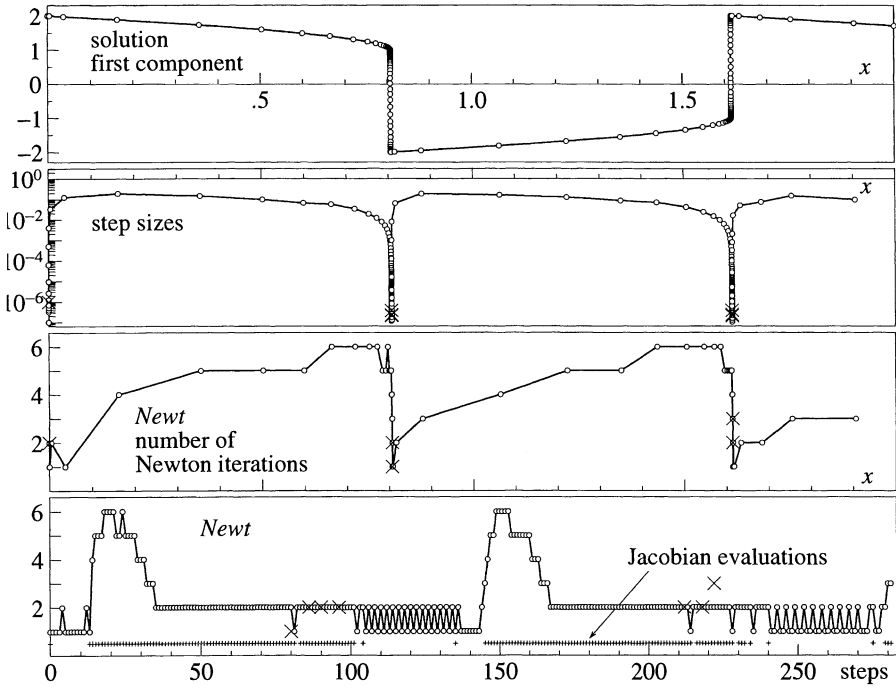


Fig. 8.1. Solution, step sizes and Newton iterations for RADAU5

In our code RADAU5 we take the minimum of the two step sizes (8.21) and (8.25). For the problem considered in Fig. 8.1, this new strategy reduces the number of rejected steps from 27 to 7.

Numerical Study of the Step-Control Mechanism. As a representative example we choose the van der Pol equation (1.5') with $\varepsilon = 10^{-6}$, initial values $y_1(0) = 2$, $y_2(0) = -0.6$ and integration interval $0 \leq x \leq 2$. Fig. 8.1 shows four pictures. The first one presents the solution $y_1(x)$ with all accepted integration steps for $Atol = Rtol = 10^{-4}$. Below this, the step sizes obtained by RADAU5 are plotted as function of x . The solid line represents the accepted steps. The rejected steps are indicated by \times 's. Observe the very small step sizes which are required in the rapid transients between the smooth parts of the solution. The lowest two pictures give the number of Newton iterations needed for solving the nonlinear system (8.2a), once as function of x , and once as function of the step-number. The last picture also indicates the steps where the Jacobian has been recomputed.

Another numerical experiment (Fig. 8.2) illustrates the quality of the error estimates. We applied the code RADAU5 with $Atol = Rtol = 10^{-4}$ and initial step size $h = 10^{-4}$ to the above problem and plotted at several chosen points of the numerical solution

- the exact local error (marked by small circles)
- the estimates (8.19) and (8.20) (marked by \diamond and \bowtie respectively)

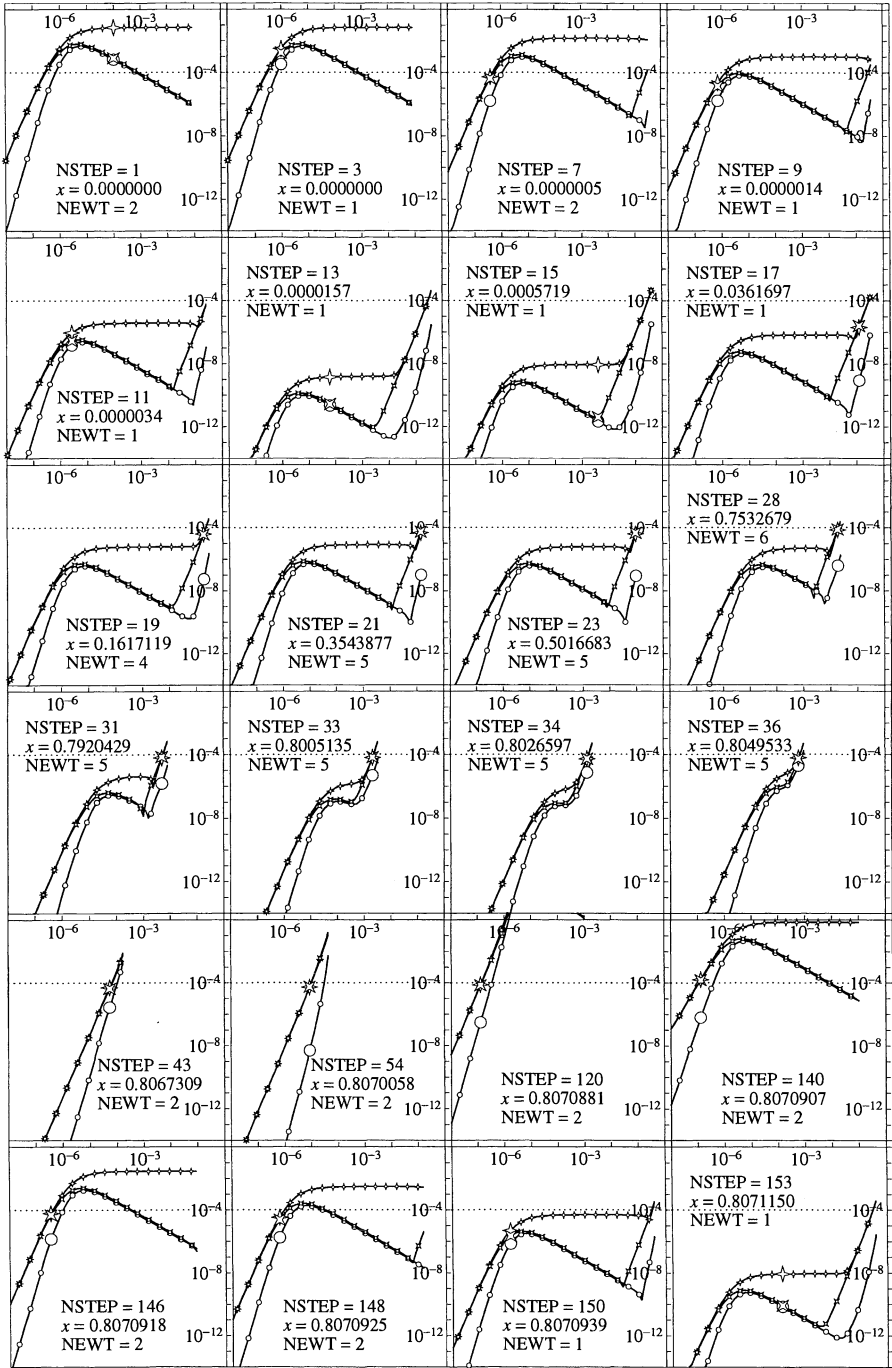


Fig. 8.2. Exact local error and the estimates (8.19) and (8.20)

as functions of h . The large symbols indicate the position of the actually used step size. *Newt* is the number of required Newton iterations.

It is interesting to note that the local error behaves like $\mathcal{O}(h^6)$ (straight line of slope 6) only for $h \leq \varepsilon$ and for large h . Between these regions, the local error *grows* like $\mathcal{O}(h^{-1})$ with decreasing h . This is the only region where the error estimate (8.20) is significantly better than (8.19). Therefore, we use the more expensive estimator (8.20) only in the first and after each rejected step. In any way, *both* error estimators are always above the actual local error, so that the code usually produces very precise results.

Implicit Differential Equations

Many applications (such as space discretizations of parabolic differential equations) often lead to systems of the form

$$My' = f(x, y), \quad y(x_0) = y_0 \quad (8.26)$$

with a constant matrix M . For such problems we formally replace all f 's by $M^{-1}f$ and multiply the resulting equations by M . Formulas (8.13) and (8.19) then have to be replaced by

$$(h^{-1}\Lambda \otimes M - I \otimes J) \Delta W^k = -h^{-1}(\Lambda \otimes M)W^k + (T^{-1} \otimes I)F((T \otimes I)W^k) \quad (8.13a)$$

$$err = ((h\gamma_0)^{-1}M - J)^{-1} (f(x_0, y_0) + (h\gamma_0)^{-1}M(e_1 z_1 + e_2 z_2 + e_3 z_3)). \quad (8.19a)$$

Here the matrix J is again an approximation to $\partial f / \partial y$. These formulas may even be applied to certain problems (8.26) with singular M (for more details see Chapters VI and VII).

Solving the linear system (8.13a) is done by a decomposition of the matrix (see (8.14), (8.14'))

$$\begin{pmatrix} \gamma M - J & 0 \\ 0 & (\alpha + i\beta)M - J \end{pmatrix}. \quad (8.27)$$

If M and J are banded or sparse, the matrices $\gamma M - J$ and $(\alpha + i\beta)M - J$ remain banded or sparse, respectively. The code RADAU5 of the appendix has options for banded structures.

An SDIRK-Code

We have also coded, using many of the above ideas, the SDIRK formula (6.16) together with the global solution (6.17). For this method also, it was again very important to replace the error estimator $y_1 - \hat{y}_1$ by (8.19).

Here, in contrast to fully implicit Runge-Kutta methods, one can treat the stages one after the other. Such a serial computation has the advantage that the information of the already computed stages can be used for a good choice of the starting values for the Newton iterations in the subsequent stages. For example, suppose that

$$\begin{aligned} z_1 &= \gamma h f(x_0 + \gamma h, y_0 + z_1) \\ z_2 &= \gamma h f(x_0 + c_2 h, y_0 + z_2) + a_{21} h f(x_0 + \gamma h, y_0 + z_1) \end{aligned}$$

are already available. Since for all i

$$z_i = c_i h f(x_0, y_0) + \left(\sum_j a_{ij} c_j \right) h^2 (f_x + f_y f)(x_0, y_0) + \mathcal{O}(h^3),$$

by solving

$$\begin{pmatrix} c_1 & c_2 \\ \sum_j a_{1j} c_j & \sum_j a_{2j} c_j \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} c_3 \\ \sum_j a_{3j} c_j \end{pmatrix}$$

one finds α_1, α_2 such that

$$\alpha_1 z_1 + \alpha_2 z_2 = z_3 + \mathcal{O}(h^3).$$

The expression $z_3^{(0)} = \alpha_1 z_1 + \alpha_2 z_2$ then serves as starting value for the computation of z_3 . In the last stage one can take \hat{y}_1 , which is then available, for starting the Newton iterations for $g_s = y_1$. The computation of z_3, z_4, y_1 , done in this way, needs few Newton iterations and a failure of convergence is usually already detected in the first stage.

However, when *parallel* processors are available, the exploitation of the triangular structure of the Runge-Kutta matrix may be less desirable. Whereas in the iteration (8.13) all s function evaluations and much of the linear algebra can be done in parallel, this is no longer possible for DIRK-methods, when z_1, \dots, z_k is used in the computations of z_{k+1} .

SIRK-Methods

The fact that singly-implicit methods have a coefficient matrix with a one-point spectrum is the key to reducing the operation count for these methods to the level which prevails in linear multistep methods.

(J.C. Butcher, K. Burrage & F.H. Chipman 1980)

In order to avoid the difficulties (in writing a Runge-Kutta code) caused by the complex eigenvalues of the Runge-Kutta matrix A , one may look for methods with real

eigenvalues, especially with a single s -fold real eigenvalue. Such methods were introduced by Nørsett (1976). Burrage (1978) provided them with error estimators, and codes in ALGOL and FORTRAN are presented in Butcher, Burrage & Chipman (1980). The basic methods for their code STRIDE are given by the following lemma.

Lemma 8.1. *For collocation methods (i.e., for Runge-Kutta methods satisfying condition $C(s)$ of Sect. IV.5), we have*

$$\det(I - zA) = (1 - \gamma z)^s, \quad (8.28)$$

if and only if

$$c_i = \gamma x_i, \quad i = 1, \dots, s, \quad (8.29)$$

where x_1, \dots, x_s are the zeros of the Laguerre polynomial $L_s(x)$ (c.f. Formula (6.11)).

Proof. The polynomial $\det(I - zA)$ is the denominator of the stability function (Formula (3.3)), so that by Theorem 3.10

$$M^{(s)}(0) + M^{(s-1)}(0)z + \dots + M(0)z^s = (1 - \gamma z)^s \quad (8.30)$$

with $M(x)$ given by (3.25). Computing $M^{(j)}(0)$ from (8.30) we obtain

$$\frac{1}{s!} \prod_{i=1}^s (x - c_i) = M(x) = \sum_{j=0}^s \binom{s}{j} (-\gamma)^{s-j} \frac{x^j}{j!} = (-\gamma)^s L_s\left(\frac{x}{\gamma}\right)$$

which leads to (8.29). □

The stability function of the method of Lemma 8.1 has been studied in Sections IV.4 (multiple real-pole approximations) and IV.6. We have further seen (Proposition 3.8) that $R(\infty) = 0$ when $x_0 + h$ is a collocation point. This means that $c_q = 1$ or $\gamma = 1/x_q$ for $q \in \{1, \dots, s\}$ where $0 < x_1 < \dots < x_s$ are the zeros of $L_s(x)$. However, if we want A -stable methods, Theorem 4.25 restricts this point to be *in the middle* (more precisely: $q = s/2$ or $s/2 + 1$ for s even, $q = (s+1)/2$ for s odd). An apparently undesirable consequence of this is that many of the collocation points lie *outside* the integration interval (for example, for $s = 5$ and $q = 3$ we have $c_1 = 0.073$, $c_2 = 0.393$, $c_3 = 1$, $c_4 = 1.970$, $c_5 = 3.515$).

Since these methods with $\gamma = 1/x_q$ are of order $p = s$ only, it is easy to embed them into a method of higher order. Burrage (1978) added a further stage

$$g_{s+1} = y_0 + h \sum_{j=1}^{s+1} a_{s+1,j} f(x_0 + c_j h, g_j)$$

where c_{s+1} and $a_{s+1,s+1}$ are arbitrary and the other $a_{s+1,j}$ are determined so that the $(s+1)$ -stage method satisfies $C(s)$ too. In order to avoid a new LU-decomposition we choose $a_{s+1,s+1} = \gamma$. The coefficient c_{s+1} is fixed arbitrarily

as $c_{s+1} = 0$. We then find a unique method

$$\hat{y}_1 = y_0 + h \sum_{j=1}^{s+1} \hat{b}_j f(x_0 + c_j h, g_j)$$

of order $s+1$ by computing the coefficients of the interpolatory quadrature rule. An explicit formula for the matrix T which transforms the Runge-Kutta matrix A to Jordan canonical form and A^{-1} to a very simple lower triangular matrix Λ is given in Exercise 1. It can be used for economically solving the linear system (8.13).

Exercises

1. (Butcher 1979). For the collocation method with c_1, \dots, c_s given by (8.29) prove that (e.g. for $s = 4$)

$$T^{-1}AT = \gamma \begin{pmatrix} 1 & & & \\ -1 & 1 & & \\ & -1 & 1 & \\ & & -1 & 1 \end{pmatrix}, \quad T^{-1}A^{-1}T = \frac{1}{\gamma} \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ 1 & 1 & 1 & \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

where the transformation T satisfies

$$T = (L_{j-1}(x_i))_{i,j=1}^s, \quad T^{-1} = \left(\frac{x_j L_{i-1}(x_j)}{s^2 L_{s-1}(x_j)^2} \right)_{i,j=1}^s$$

and $L_{j-1}(x)$ are the Laguerre polynomials.

Hint. Use the identities

$$L'_n(x) = L'_{n-1}(x) - L_{n-1}(x), \quad L_n(x) = L_{n-1}(x) + \frac{x}{n} L'_n(x)$$

and the Christoffel-Darboux formula

$$\sum_{j=0}^n L_j(x) L_j(y) = \frac{n+1}{y-x} (L_{n+1}(x) L_n(y) - L_{n+1}(y) L_n(x))$$

which, in the limit $y \rightarrow x$, becomes

$$\sum_{j=0}^n (L_j(x))^2 = (n+1) (L_{n+1}(x) L'_n(x) - L'_{n+1}(x) L_n(x)).$$

IV.9 Extrapolation Methods

It seems that a suitable version of an IEM (implicit extrapolation method) which takes care of these difficulties may become a very strong competitor to any of the general discretization methods for stiff systems presently known.

(the very last sentence of Stetter's book, 1973)

Extrapolation of explicit methods is an interesting approach to solving nonstiff differential equations (see Sect. II.9). Here we show to what extent the idea of extrapolation can also be used for stiff problems. We shall use the results of Sect. II.8 for the existence of asymptotic expansions and apply them to the study of those implicit and linearly implicit methods, which seem to be most suitable for the computation of stiff differential equations. Our theory here is restricted to classical $h \rightarrow 0$ order, the study of stability domains and A -stability.

A big difficulty, however, is the fact that the coefficients and remainders of the asymptotic expansion can explode with increasing stiffness and the h -interval, for which the expansion is meaningful, may tend to zero. Bounds on the remainder which hold uniformly for a class of arbitrarily stiff problems, will be discussed later in Sect. VI.5.

Extrapolation of Symmetric Methods

It is most natural to look first for symmetric one-step methods as the basic integration scheme. Promising candidates are the trapezoidal rule

$$y_{i+1} = y_i + \frac{h}{2} \left(f(x_i, y_i) + f(x_{i+1}, y_{i+1}) \right) \quad (9.1)$$

and the implicit mid-point rule

$$y_{i+1} = y_i + hf \left(x_i + \frac{h}{2}, \frac{1}{2} (y_{i+1} + y_i) \right). \quad (9.2)$$

We take some step-number sequence $n_1 < n_2 < n_3 < \dots$, set $h_j = H/n_j$ and define

$$T_{j1} = y_{h_j}(x_0 + H), \quad (9.3)$$

the numerical solution obtained by performing n_j steps with step size h_j . As described in Sect. II.9 we extrapolate these values according to

$$T_{j,k+1} = T_{j,k} + \frac{T_{j,k} - T_{j-1,k}}{(n_j/n_{j-k})^2 - 1}. \quad (9.4)$$

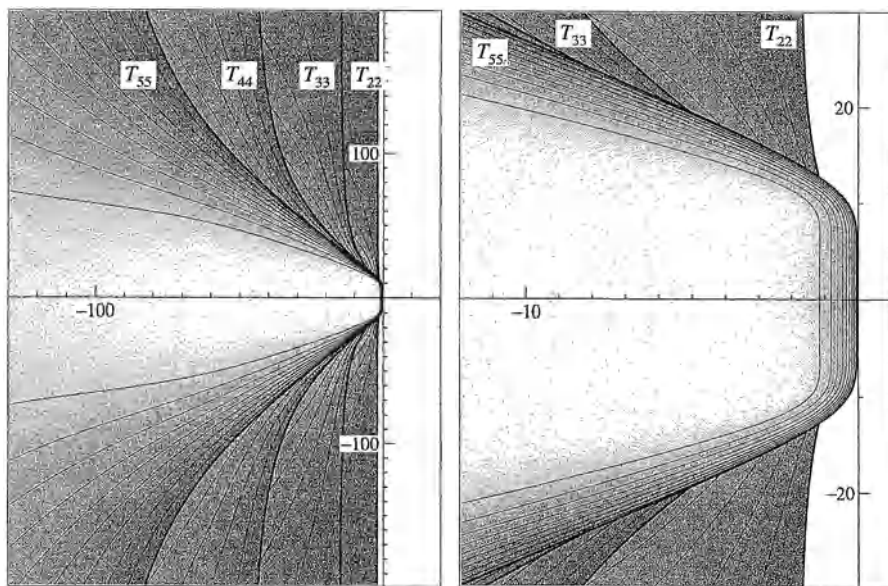


Fig. 9.1. Stability domains for the extrapolated trapezoidal rule

This provides an extrapolation tableau

$$\begin{array}{cccc}
 T_{11} & & & \\
 T_{21} & T_{22} & & \\
 T_{31} & T_{32} & T_{33} & \\
 \vdots & \vdots & \vdots & \ddots,
 \end{array} \tag{9.5}$$

all entries of which represent diagonally implicit Runge-Kutta methods (see Exercise 1). Due to the symmetry of the basic schemes (9.1) and (9.2), T_{jk} is a DIRK-method of order $2k$. In order to study the stability properties of these methods, we apply them to the test equation $y' = \lambda y$. For both methods, (9.1) and (9.2), we obtain

$$y_{i+1} = \frac{1 + \frac{h\lambda}{2}}{1 - \frac{h\lambda}{2}} y_i$$

so that the stability function $R_{jk}(z)$ of the method T_{jk} is given recursively by ($z = H\lambda$)

$$R_{j1}(z) = \left(\frac{1 + \frac{z}{2n_j}}{1 - \frac{z}{2n_j}} \right)^{n_j}, \tag{9.6a}$$

$$R_{j,k+1}(z) = R_{j,k}(z) + \frac{R_{j,k}(z) - R_{j-1,k}(z)}{(n_j/n_{j-k})^2 - 1}. \tag{9.6b}$$

Already Dahlquist (1963) noticed that for $n_1 = 1$ and $n_2 = 2$ we have

$$R_{22}(z) = \frac{1}{3} \left(4 \left(\frac{1 + \frac{z}{4}}{1 - \frac{z}{4}} \right)^2 - \left(\frac{1 + \frac{z}{2}}{1 - \frac{z}{2}} \right) \right) \rightarrow \frac{5}{3} > 1 \quad \text{for } z \rightarrow \infty, \quad (9.7)$$

an undesirable property when solving stiff problems. Stetter (1973) proposed taking only even or only odd numbers in the step-number sequence $\{n_j\}$. Then, all stability functions of the extrapolation tableau tend for $z \rightarrow \infty$ to 1 or -1 , respectively. But even in this situation extrapolation immediately destroys the A -stability of the underlying scheme (Exercise 2). Fig. 9.1 shows the stability domains $\{z; |R_{kk}(z)| \leq 1\}$ for the sequence $\{1, 3, 5, 7, 9, \dots\}$.

Smoothing

Some numerical examples reveal the power of the smoothing combined with extrapolation. (B. Lindberg 1971)

Another possibility to overcome the difficulty encountered in (9.7) is smoothing (Lindberg 1971). The idea is to replace the definition (9.3) by Gragg's smoothing step

$$\hat{T}_{j1} = S_{h_j}(x_0 + H), \quad (9.8)$$

$$S_h(x) = \frac{1}{4}(y_h(x-h) + 2y_h(x) + y_h(x+h)). \quad (9.9)$$

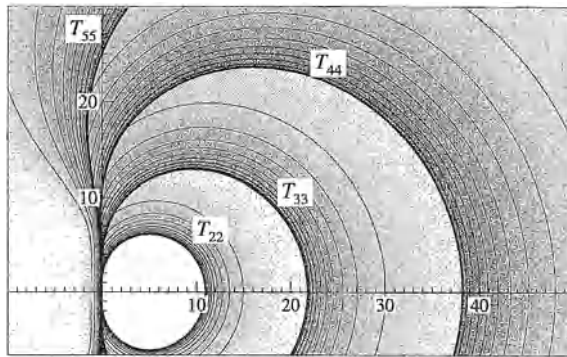
With $y_h(x)$, $S_h(x)$ also possesses an asymptotic expansion in even powers of h . Therefore, extrapolation according to (9.4) is justified. For the stability function of \hat{T}_{j1} we now obtain

$$\begin{aligned} \hat{R}_{j1}(z) &= \frac{1}{4} \left\{ \left(\frac{1 + \frac{z}{2n_j}}{1 - \frac{z}{2n_j}} \right)^{n_j-1} + 2 \left(\frac{1 + \frac{z}{2n_j}}{1 - \frac{z}{2n_j}} \right)^{n_j} + \left(\frac{1 + \frac{z}{2n_j}}{1 - \frac{z}{2n_j}} \right)^{n_j+1} \right\} \\ &= \frac{1}{\left(1 - \frac{z}{2n_j}\right)^2} \left(\frac{1 + \frac{z}{2n_j}}{1 - \frac{z}{2n_j}} \right)^{n_j-1} \end{aligned} \quad (9.10)$$

which is an L -stable approximation to the exponential function. The stability functions $\hat{R}_{jk}(z)$ (obtained from (9.6b)) all satisfy $\hat{R}_{jk}(z) = \mathcal{O}(z^{-2})$ for $z \rightarrow \infty$. For the step-number sequence

$$\{n_j\} = \{1, 2, 3, 4, 5, 6, 7, \dots\} \quad (9.11)$$

the stability domains of $\hat{R}_{kk}(z)$ are plotted in Fig. 9.2.

Fig. 9.2. Stability domains of $\hat{R}_{kk}(z)$

The Linearly Implicit Mid-Point Rule

Extrapolation codes based on fully implicit methods are difficult to implement efficiently. After extensive numerical computations, G. Bader and P. Deuflhard (1983) found that a linearly implicit (Rosenbrock-type) extension of the GBS method of Sect. II.9 gave promising results for stiff equations. This method is based on a two-step algorithm, since one-step Rosenbrock methods (7.4) cannot be symmetric for nonlinear differential equations.

The motivation for the Bader & Deuflhard method is based on Lawson's transformation (Lawson 1967)

$$y(x) = e^{Jx} \cdot c(x), \quad (9.12)$$

where it is hoped that the matrix $J \approx f'(y)$ will neutralize the stiffness. Differentiation gives

$$c' = e^{-Jx} \cdot g(x, e^{Jx}c) \quad \text{with} \quad g(x, y) = f(x, y) - Jy. \quad (9.13)$$

We now solve (9.13) by the Gragg algorithm (II.9.13b)

$$c_{i+1} = c_{i-1} + 2he^{-Jx_i} \cdot g(x_i, e^{Jx_i}c_i)$$

and obtain by back-substitution of (9.12)

$$e^{-hJ}y_{i+1} = e^{hJ}y_{i-1} + 2hg(x_i, y_i). \quad (9.14)$$

For evident reasons of computational ease we now replace $e^{\pm hJ}$ by the approximations $I \pm hJ$ and obtain, adding an appropriate starting and final smoothing step,

$$(I - hJ)y_1 = y_0 + hg(x_0, y_0) \quad (9.15a)$$

$$(I - hJ)y_{i+1} = (I + hJ)y_{i-1} + 2hg(x_i, y_i) \quad (9.15b)$$

$$S_h(x) = \frac{1}{2}(y_{2m-1} + y_{2m+1}) \quad \text{where } x = x_0 + 2mh. \quad (9.15c)$$

Substituting finally g from (9.13), we arrive at (with $x = x_0 + 2mh$, $x_i = x_0 + ih$)

$$(I - hJ)(y_1 - y_0) = hf(x_0, y_0) \quad (9.16a)$$

$$(I - hJ)(y_{i+1} - y_i) = -(I + hJ)(y_i - y_{i-1}) + 2hf(x_i, y_i) \quad (9.16b)$$

$$S_h(x) = \frac{1}{2}(y_{2m-1} + y_{2m+1}) \quad (9.16c)$$

where J stands for some approximation to the Jacobian $\frac{\partial f}{\partial y}(x_0, y_0)$. Putting $J = 0$, Formulas (9.16a) and (9.16b) become equivalent to those of the GBS method. The scheme (9.16b) is the linearly implicit (or semi-implicit) mid-point rule, Formula (9.16a) the linearly implicit Euler method.

Theorem 9.1 (Bader & Deuffhard 1983). *Let $f(x, y)$ be sufficiently often differentiable and let J be an arbitrary matrix; then the numerical solution defined by (9.16a,b,c) possesses an asymptotic expansion of the form*

$$y(x) - S_h(x) = \sum_{j=1}^l e_j(x)h^{2j} + h^{2l+2}C(x, h) \quad (9.17)$$

where $C(x, h)$ is bounded for $x_0 \leq x \leq \bar{x}$ and $0 \leq h \leq h_0$. For $J \neq 0$ we have in general $e_j(x_0) \neq 0$.

Proof. As in Stetter's proof for the GBS algorithm we introduce the variables

$$\begin{aligned} h^* &= 2h, \quad x_k^* = x_0 + kh^*, \quad u_0 = v_0 = y_0, \quad u_k = y_{2k}, \\ v_k &= (I - hJ)y_{2k+1} + hJy_{2k} - hf(x_{2k}, y_{2k}) \\ &= (I + hJ)y_{2k-1} - hJy_{2k} + hf(x_{2k}, y_{2k}). \end{aligned} \quad (9.18)$$

Method (9.16a,b) can then be rewritten as

$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} u_k \\ v_k \end{pmatrix} + h^* \begin{pmatrix} f(x_k^* + \frac{h^*}{2}, y_{2k+1}) - Jy_{2k+1} + J(\frac{u_{k+1} + u_k}{2}) \\ \frac{1}{2} \left(f(x_k^* + h^*, u_{k+1}) + f(x_k^*, u_k) \right) + Jy_{2k+1} - J(\frac{u_{k+1} + u_k}{2}) \end{pmatrix} \quad (9.19)$$

where, from (9.18), we obtain the symmetric representation

$$y_{2k+1} = \frac{v_{k+1} + v_k}{2} + h^* J \left(\frac{u_{k+1} - u_k}{4} \right) - \frac{h^*}{4} \left(f(x_{k+1}^*, u_{k+1}) - f(x_k^*, u_k) \right).$$

The symmetry of (9.19) is illustrated in Fig. 9.3 and can be checked analytically by exchanging $u_{k+1} \leftrightarrow u_k$, $v_{k+1} \leftrightarrow v_k$, $h^* \leftrightarrow -h^*$, and $x_k^* \leftrightarrow x_k^* + h^*$. Method (9.19) is consistent with the differential equation

$$\begin{aligned} u' &= f(x, v) - J(v - u), & u(x_0) &= y_0 \\ v' &= f(x, u) + J(v - u), & v(x_0) &= y_0 \end{aligned}$$

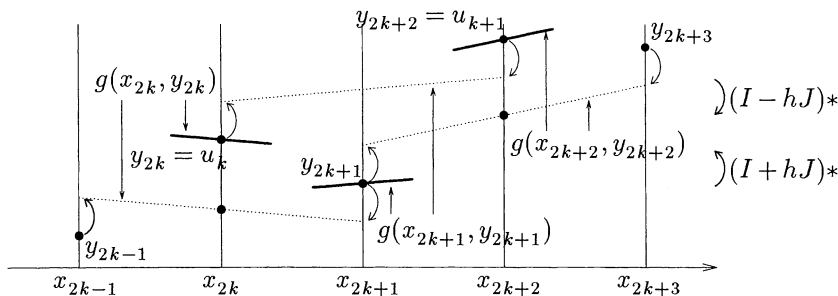


Fig. 9.3. Symmetry of Method (9.19) (see (9.16b))

whose exact solution is $u(x) = v(x) = y(x)$, where $y(x)$ is the solution of the original equation $y' = f(x, y)$. Applying Theorem II.8.10 we obtain

$$\begin{aligned} y(x) - u_{h*}(x) &= \sum_{j=1}^l a_j(x) h^{2j} + h^{2l+2} A(x, h) \\ y(x) - v_{h*}(x) &= \sum_{j=1}^l b_j(x) h^{2j} + h^{2l+2} B(x, h) \end{aligned} \quad (9.20)$$

with $a_j(x_0) = b_j(x_0) = 0$. With the help of Formulas (9.18) we can express the numerical solution (9.16c) in terms of u_m and v_m as follows:

$$\frac{1}{2}(y_{2m+1} + y_{2m-1}) = (I - h^2 J^2)^{-1} \left(v_m + h^2 J(f(x_{2m}, u_m) - J u_m) \right),$$

and we obtain for $x = x_0 + 2mh$,

$$\begin{aligned} y(x) - S_h(x) &= (I - h^2 J^2)^{-1} \left(y(x) - v_{h*}(x) \right. \\ &\quad \left. - h^2 J \left(f(x, u_{h*}(x)) + J(y(x) - u_{h*}(x)) \right) \right). \end{aligned}$$

Inserting the expansions (9.20) we find (9.17). \square

As an application of this theorem we obtain an interesting theoretical result on the existence of W -methods (7.4) (with inexact Jacobian). We saw in Volume I (Exercise 1 of Sect. II.9 and Theorem II.9.4) that the $T_{j,k}$ of the extrapolated GBS method represent explicit Runge-Kutta methods. By analogy, it is not difficult to guess that the $T_{j,k}$ for the above linearly implicit midpoint rule represent W -methods (more details in Exercise 3) and we have the following existence result for such methods.

Theorem 9.2. *For p even, there exists a W -method (7.4) of order p with $s = p(p+2)/4$ stages.*

Proof. It follows from (9.20) that for $x = x_0 + 2mh$ the numerical solution $y_h(x) = y_{2m}$ possesses an h^2 -expansion of the form (9.17) with $e_j(x_0) = 0$. Therefore, extrapolation yields W -methods of order $2k$ (in the k -th column). The result follows by taking $\{n_j\} = \{2, 4, 6, 8, 10, 12, \dots\}$ and counting the number of necessary function evaluations. \square

Table 9.1. $A(\alpha)$ -stability of extrapolated linearly implicit mid-point rule

90°						
90°	90°					
90°	90°	90°				
90°	89.34°	87.55°	87.34°			
90°	88.80°	86.87°	86.10°	86.02°		
90°	88.49°	87.30°	86.61°	86.36°	86.33°	
90°	88.43°	87.42°	87.00°	86.78°	86.70°	86.69°

For a stability analysis we apply the method (9.16) with $J = \lambda$ to the test equation $y' = \lambda y$. In this case Formula (9.16b) reduces to

$$y_{i+1} = \frac{1 + h\lambda}{1 - h\lambda} y_{i-1}$$

and the numerical result is given by

$$S_h(x_0 + 2mh) = \frac{1}{(1 - h\lambda)^2} \left(\frac{1 + h\lambda}{1 - h\lambda} \right)^{m-1} y_0, \quad (9.21)$$

exactly the same as that obtained from the trapezoidal rule with smoothing (see Formula (9.10)). We next have to choose a step-number sequence $\{n_j\}$. Clearly, $n_j = 2m_j$ must be even. Bader & Deuffhard (1983) proposed taking only odd numbers m_j , since then $S_h(x_0 + 2m_j h)$ in (9.21) has the same sign as the exact solution $e^{\lambda 2m_j h} y_0$ for all real $h\lambda \leq 0$. Consequently they were led to

$$\{n_j\} = \{2, 6, 10, 14, 22, 34, 50, \dots\}. \quad (9.22)$$

Putting $T_{j1} = S_{h_j}(x_0 + H)$ with $h_j = H/n_j$ and defining T_{jk} by (9.4) we obtain a tableau of W -methods (7.4) (Exercise 3). By Theorem 9.1 the k -th column of this tableau represents methods of order $2k - 1$ independent of the choice of J (the methods are not of order $2k$, since $e_l(x_0) \neq 0$ in (9.17)). The stability function of T_{j1} is given by

$$R_{j1}(z) = \frac{1}{\left(1 - \frac{z}{n_j}\right)^2} \left(\frac{1 + \frac{z}{n_j}}{1 - \frac{z}{n_j}} \right)^{n_j/2-1} \quad (9.23)$$

and those of T_{jk} can be computed with the recursion (9.6b). An investigation of the E -polynomial (3.8) for these rational functions shows that not only T_{j1} , but

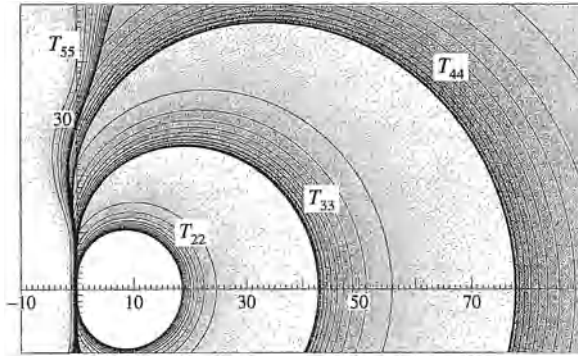


Fig. 9.4. Stability domains of extrapolated linearly implicit mid-point rule

also T_{22} , T_{32} and T_{33} are A -stable (Hairer, Bader & Lubich 1982). The angle of $A(\alpha)$ -stability for some further elements in the extrapolation tableau are listed in Table 9.1. Stability domains of T_{kk} for $k = 2, 3, 4, 5, 6$ are plotted in Fig. 9.4.

Implicit and Linearly Implicit Euler Method

Why not consider also non-symmetric methods as basic integration schemes? Deuflhard (1985) reports on experiments with extrapolation of the implicit Euler method

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}) \quad (9.24)$$

and of the linearly implicit Euler method

$$(I - hJ)(y_{i+1} - y_i) = hf(x_i, y_i), \quad (9.25)$$

where, again, J is an approximation to $\frac{\partial f}{\partial y}(x_0, y_0)$. These methods are not symmetric and have only a h -expansion of their global error. We therefore have to extrapolate the numerical solutions at $x_0 + H$ according to

$$T_{j,k+1} = T_{j,k} + \frac{T_{j,k} - T_{j-1,k}}{(n_j/n_{j-k}) - 1}, \quad (9.26)$$

so that T_{jk} represents a method of order k .

For both basic methods, (9.24) and (9.25), the stability function of T_{jk} is the same and defined recursively by

$$R_{j1}(z) = \left(1 - \frac{z}{n_j}\right)^{-n_j} \quad (9.27a)$$

$$R_{j,k+1}(z) = R_{j,k}(z) + \frac{R_{j,k}(z) - R_{j-1,k}(z)}{(n_j/n_{j-k}) - 1}. \quad (9.27b)$$

Taking the step-number sequence

$$\{n_j\} = \{1, 2, 3, 4, 5, 6, 7, \dots\} \quad (9.28)$$

we have plotted in Fig. 9.5 the stability domains of $R_{kk}(z)$ (left picture) and $R_{k,k-1}(z)$ (right picture). All these methods are seen to be $A(\alpha)$ -stable with α close to 90° . The values of α (computed numerically) for $R_{jk}(z)$ with $j \leq 8$ are given in Table 9.2.

We shall see in the chapter on differential algebraic systems that it is preferable to use the first subdiagonal of the extrapolation tableau resulting from (9.28). This is equivalent to the use of the step number sequence $\{n_i\} = \{2, 3, 4, 5, \dots\}$. Also an effective construction of a *dense output* can best be motivated in the setting of differential-algebraic equations (Sect. VI.5).

Table 9.2. $A(\alpha)$ -stability of extrapolated Euler

90°							
90°	90°						
90°	90°	89.85°					
90°	90°	89.90°	89.77°				
90°	90°	89.93°	89.84°	89.77°			
90°	90°	89.95°	89.88°	89.82°	89.78°		
90°	90°	89.96°	89.91°	89.86°	89.82°	89.80°	
90°	90°	89.97°	89.93°	89.89°	89.85°	89.83°	89.81°

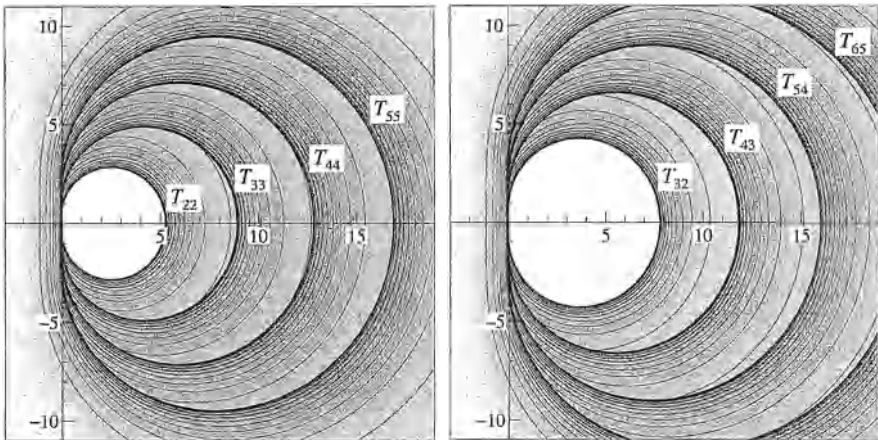


Fig. 9.5. Stability domains of extrapolated Euler

Implementation

Extrapolation methods based on implicit discretizations are in general less efficient than those based on linearly implicit discretizations. The reason is that the arising nonlinear systems have to be solved very accurately, so that the asymptotic expan-

sion of the error is not destroyed. The first successful extrapolation code for stiff differential equations is METAN1 of Bader & Deuffhard (1983), which implements the linearly implicit mid-point rule (9.16). In fact, Formula (9.16b) is replaced by the equivalent formulation

$$\Delta y_i = \Delta y_{i-1} + 2(I - hJ)^{-1} \left(hf(x_i, y_i) - \Delta y_{i-1} \right), \quad \Delta y_i = y_{i+1} - y_i \quad (9.29)$$

which avoids a matrix-vector multiplication. The step size and order selection of this code is described in Deuffhard (1983). Modifications in the control of step size and order are proposed by Shampine (1987). We have implemented the following two extrapolation codes (see Appendix):

SODEX is based on the linearly implicit mid-point rule (9.16), uses the step-number sequence (9.22) and is mathematically equivalent to METAN1. The step size and order selection in SODEX is with some minor changes that of the non-stiff code ODEX of Sect. II.9. We just mention that in the formula for the work per unit step (II.9.26) the number A_k is augmented by the dimension of the differential equation in order to take into account the Jacobian evaluation.

SEULEX is an implementation of the linearly implicit Euler method (9.25) using the step-number sequence $\{2, 3, 4, 5, 6, 7, \dots\}$ (other sequences can be chosen as internal options). The step size and order selection is that of SODEX. The original code (EULSIM, first discussed by Deuffhard 1985) uses the same numerical method, but a different implementation.

Neither code can solve the van der Pol equation problem in a straightforward way because of overflow ...

(L.F. Shampine 1987)

A big difficulty in the implementation of extrapolation methods is the use of “large” step sizes. During the computation of T_{j1} one may easily get into trouble with exponential overflow when evaluating the right-hand side of the differential equation. As a remedy we propose the following strategies:

- a) In establishing the extrapolation tableau we compare the estimated error $err_j = \|T_{j,j-1} - T_{jj}\|$ with the preceding one. Whenever $err_j \geq err_{j-1}$ for some $j \geq 3$ we restart the computation of the step with a smaller H , say, $H = 0.5 \cdot H$.
- b) In order to be able to interrupt the computations already after the first f -evaluations, we require that the step sizes $h = H/n_i$ (for $i = 1$ and $i = 2$) be small enough so that a simplified Newton iteration applied to the implicit Euler method $y = y_0 + hf(x, y)$, $x = x_0 + h$ would converge (“stability check”, an idea of Deuffhard). The first two iterations read

$$\begin{aligned} (I - hJ)\Delta_0 &= hf(x_0, y_0), & y^{(1)} &= y_0 + \Delta_0 \\ (I - hJ)\Delta_1 &= hf(x_0 + h, y^{(1)}) - \Delta_0. \end{aligned} \quad (9.30)$$

The computations for the step are restarted with a smaller H , if $\|\Delta_1\| \geq \|\Delta_0\|$

(divergence of the iteration). Observe that for both methods, (9.16) and (9.25), no additional function evaluations are necessary. For the linearly implicit mid-point rule we have the simple relations $\Delta_0 = \Delta y_0$, $\Delta_1 = \frac{1}{2}(\Delta y_1 - \Delta y_0)$ (see (9.29)).

Non-Autonomous Differential Equations. Given a non-autonomous differential equation $y' = f(x, y)$, one has several possibilities to apply the above extrapolation algorithms:

- i) apply the Formula (9.16) or (9.25) directly (this is justified, since all asymptotic expansions hold for general non-autonomous problems);
- ii) transform the differential equation into an autonomous system by adding $x' = 1$ and then apply the algorithm. This yields

$$(I - hJ)(y_{i+1} - y_i) = hf(x_i, y_i) + h^2 \frac{\partial f}{\partial x}(x_0, y_0) \quad (9.31)$$

for the linearly implicit Euler method (the derivative $\frac{\partial f}{\partial x}(x_0, y_0)$ can also be replaced by some approximation). For the linearly implicit mid-point rule, (9.16a) has to be replaced by (9.31) with $i = 0$, the remaining two formulas (9.16b) and (9.16c) are not changed.

- iii) apply one simplified Newton iteration to the implicit Euler discretization (9.24). This gives

$$(I - hJ)(y_{i+1} - y_i) = hf(x_{i+1}, y_i). \quad (9.32)$$

The use of this formula avoids the computation of the derivative $\partial f / \partial x$, but requires one additional function evaluation for each T_{j1} . In the case of the linearly implicit mid-point rule the replacement of (9.16a) by (9.32) would destroy symmetry and the expansions in h^2 .

A theoretical study of the three different approaches for the linearly implicit Euler method applied to the Prothero-Robinson equation (see Exercise 4 below) indicates that the third approach is preferable. More theoretical insight into this question will be obtained from the study of singular perturbation problems (Chapter VI).

Implicit Differential Equations. Our codes in the appendix are written for problems of the form

$$My' = f(x, y) \quad (9.33)$$

where M is a constant square matrix. The necessary modifications in the basic formulas are obtained, as usual, by replacing all f 's and J 's by $M^{-1}f$ and $M^{-1}J$, and premultiplying by M . The linearly implicit Euler method then reads

$$(M - hJ)(y_{i+1} - y_i) = hf(x_i, y_i) \quad (9.34)$$

and the linearly implicit mid-point rule becomes, with $\Delta y_i = y_{i+1} - y_i$,

$$\Delta y_i = \Delta y_{i-1} + 2(M - hJ)^{-1} \left(hf(x_i, y_i) - M \Delta y_{i-1} \right). \quad (9.35)$$

Exercises

- Consider the implicit mid-point rule (9.2) as basic integration scheme and define T_{jk} by (9.3) and (9.4).
 - Prove that T_{jk} represents a DIRK-method of order $p = 2k$ with $s = n_1 + n_2 + \dots + n_j$ stages.
 - \hat{T}_{jk} , defined by (9.8) and (9.4), is equivalent to a DIRK-method of order $p = 2k - 1$ only.
- Let $R_{jk}(z)$ be given by (9.6) and assume that the step-number sequence consists of even numbers only. Prove that $R_{j2}(z)$ cannot be A -stable. More precisely, show that at most a finite number of points of the imaginary axis can lie in the stability domain of $R_{j2}(z)$ (interpret Fig. 9.6).

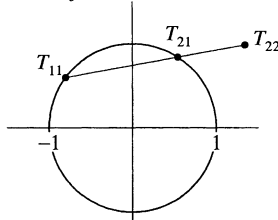


Fig. 9.6. How extrapolation destroys A -stability

- Prove that $S_h(x)$, defined by (9.16), is the numerical result of the $(2n + 1)$ -stage W -method (7.4) with the following coefficients ($n = 2m$):

$$\alpha_{ij} = \begin{cases} 1/n & \text{if } j = 1 \text{ and } i \text{ even,} \\ 2/n & \text{if } 1 < j < i \text{ and } i - j \text{ odd,} \\ 0 & \text{else.} \end{cases}$$

$$\gamma_{ij} = \begin{cases} (-1)^{i-j}/n & \text{if } j = 1 \text{ or } j = i, \\ 2(-1)^{i-j}/n & \text{if } 1 < j < i. \end{cases}$$

$$b_i = \alpha_{n+1,i} + \gamma_{n+1,i} \quad \text{for all } i.$$

- Apply the three different versions of the linearly implicit Euler method (9.25), (9.31) and (9.32) to the problem $y' = \lambda(y - \varphi(x)) + \varphi'(x)$. Prove that the errors $e_i = y_i - \varphi(x_i)$ satisfy $e_{i+1} = (1 - h\lambda)^{-1} e_i + \delta_h(x_i)$, where for $h \rightarrow 0$ and $h\lambda \rightarrow \infty$,

$$\delta_h(x) = -h\varphi'(x) + \mathcal{O}(h^2) + \mathcal{O}(\lambda^{-1}),$$

$$\delta_h(x) = -\frac{h^2}{2}\varphi''(x) + (1 - h\lambda)^{-1}h^2\lambda(\varphi'(x) - \varphi'(x_0)) + \mathcal{O}(h^3) + \mathcal{O}(h\lambda^{-1}),$$

$$\delta_h(x) = (1 - h\lambda)^{-1}\left(\frac{h^2}{2}\varphi''(x) + \mathcal{O}(h^3)\right),$$

respectively.

IV.10 Numerical Experiments

Theory without practice cannot survive and dies as quickly as it lives.
(Leonardo da Vinci
1452-1519, cited from M. Kline, *Math. Thought* 1972, p. 224)

Sine experientia nihil sufficienter scribere potest (Without experience it is not possible to know anything adequately).
(Inscription overlooking Botanic Garden, Oxford; found in *The Latin Citation Calendar*, Oxford 1996)

After having seen so many different methods and ideas in the foregoing sections, it is legitimate to study how all these theoretical properties pay off in numerical efficiency.

The Codes Used

We compared the following codes, some of which are described in the Appendix:

RADAU5 and **SDIRK4** are implicit Runge-Kutta codes. The first one is based on the Radau IIA method with $s = 3$ of order 5 (Table 5.6), whereas the second one is based on the SDIRK method (6.16) of order 4. Both methods are L -stable. Details of their implementation are given in Sect. IV.8.

RODAS and **ROS4** are Rosenbrock codes of order 4 with an embedded 3rd order error estimator. **ROS4** implements the methods of Table 7.2. A switch allows one to choose between the different coefficient sets. The underlying method of **RODAS** satisfies additional order conditions for differential-algebraic equations (see Sect. VI.4 below), but requires a little more work per step. **RODAS5** is an extension of **RODAS** to order 5. Its coefficients are constructed by Di Marzo (1992).

SEULEX and **SODEX** are extrapolation codes. They implement the (Stiff) linearly implicit **EULER EXtrapolation** method (9.32) and the extrapolation algorithm based on the linearly implicit mid-point rule (method (9.16) of Bader & Deuffhard 1983), respectively. Both methods are discussed in Sect. IV.9.

In the numerical experiments of this section we have also included the results of **LSODE** (a BDF code of Hindmarsh 1980). It is a representative of the class of multistep methods to be described in Chapter V.

Many of the treated examples are very stiff and *explicit* methods would require hours to compute the solution. On some examples, however, it was also interesting to see their performance, especially for the methods with extended region of stability (e.g., the Runge-Kutta-Chebyshev code **RKC** of Sommeijer (1991), explained in Sect. IV.2), as well as for a standard explicit Runge-Kutta code, such as **DOPRI5** of Volume I.

Twelve Test Problems

Man hüte sich, auf Grund einzelner Beispiele allgemeine Schlüsse über den Wert oder Unwert einer Methode zu ziehen. Dazu gehört sehr viel Erfahrung. (L. Collatz 1950)

The first extensive numerical comparisons for stiff equations were made by Enright, Hull & Lindberg (1975). Their STIFF-DETEST set of problems has become a veritable “must” for generations of software writers (see also the critical remarks of Shampine 1981). Several additional test problems, usually from chemical kinetics, have been proposed by Enright & Hull (1976). An interesting review article containing also problems of large dimension is due to Byrne & Hindmarsh (1987).

The problems chosen here for our tests are the following:

VDPOL — the van der Pol oscillator (see (1.5') and Fig. 8.1)

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= ((1 - y_1^2)y_2 - y_1)/\varepsilon, \quad \varepsilon = 10^{-6} \\ y_1(0) &= 2, \quad y_2(0) = 0; \quad x_{\text{out}} = 1, 2, 3, 4, \dots, 11. \end{aligned} \quad (10.1)$$

ROBER — the reaction of Robertson (1966) (see (1.3) and (1.4))

$$\begin{aligned} y_1' &= -0.04y_1 + 10^4 y_2 y_3 & y_1(0) &= 1 \\ y_2' &= 0.04y_1 - 10^4 y_2 y_3 - 3 \cdot 10^7 y_2^2 & y_2(0) &= 0 \\ y_3' &= 3 \cdot 10^7 y_2^2 & y_3(0) &= 0, \end{aligned} \quad (10.2)$$

one of the most prominent examples of the “stiff” literature. It was usually treated on the interval $0 \leq x \leq 40$, until Hindmarsh discovered that many codes fail if x becomes very large (10^{11} say). The reason is that whenever the numerical solution of y_2 accidentally becomes negative, it then tends to $-\infty$ and the run ends by overflow. We have therefore chosen $x_{\text{out}} = 1, 10, 10^2, 10^3, \dots, 10^{11}$.

OREGO — the Oregonator, the famous model with a periodic solution describing the Belusov-Zhabotinskii reaction (Field & Noyes 1974, see also Enright & Hull 1976)

$$\begin{aligned} y_1' &= 77.27(y_2 + y_1(1 - 8.375 \cdot 10^{-6} y_1 - y_2)) \\ y_2' &= \frac{1}{77.27}(y_3 - (1 + y_1)y_2) \\ y_3' &= 0.161(y_1 - y_3) \end{aligned} \quad (10.3)$$

$$y_1(0) = 1, \quad y_2(0) = 2, \quad y_3(0) = 3, \quad x_{\text{out}} = 30, 60, 90, \dots, 360.$$

For pictures see Volume I, p. 119.

HIRES — this chemical reaction involving eight reactants was proposed by Schäfer (1975) to explain “the growth and differentiation of plant tissue independent of

photosynthesis at high levels of irradiance by light". It has been promoted as a test example by Gottwald (1977). The corresponding equations are

$$\begin{aligned}
 y_1' &= -1.71 \cdot y_1 + 0.43 \cdot y_2 + 8.32 \cdot y_3 + 0.0007 \\
 y_2' &= 1.71 \cdot y_1 - 8.75 \cdot y_2 \\
 y_3' &= -10.03 \cdot y_3 + 0.43 \cdot y_4 + 0.035 \cdot y_5 \\
 y_4' &= 8.32 \cdot y_2 + 1.71 \cdot y_3 - 1.12 \cdot y_4 \\
 y_5' &= -1.745 \cdot y_5 + 0.43 \cdot y_6 + 0.43 \cdot y_7 \\
 y_6' &= -280 \cdot y_6 y_8 + 0.69 \cdot y_4 + 1.71 \cdot y_5 - 0.43 \cdot y_6 + 0.69 \cdot y_7 \\
 y_7' &= 280 \cdot y_6 y_8 - 1.81 \cdot y_7 \\
 y_8' &= -y_7'
 \end{aligned} \tag{10.4}$$

$$y_1(0) = 1, \quad y_2(0) = y_3(0) = \dots = y_7(0) = 0, \quad y_8(0) = 0.0057$$

and chosen output values are $x_{\text{out}} = 321.8122$ and 421.8122 .

E5 — is another chemical reaction problem, called "E5" in the collection by Enright, Hull & Lindberg (1975). It is given by

$$\begin{aligned}
 y_1' &= -Ay_1 - By_1y_3 & y_1(0) &= 1.76 \cdot 10^{-3} \\
 y_2' &= Ay_1 & y_2(0) &= 0 \\
 y_3' &= Ay_1 - By_1y_3 - MCy_2y_3 + Cy_4 & y_3(0) &= 0 \\
 y_4' &= By_1y_3 & y_4(0) &= 0,
 \end{aligned} \tag{10.5}$$

where $A = 7.89 \cdot 10^{-10}$, $B = 1.1 \cdot 10^7$, $C = 1.13 \cdot 10^3$, and $M = 10^6$. As we can see from Fig. 10.1 the variables are badly scaled ($y_1 \approx 10^{-3}$ at the beginning, all other components do not exceed the value $1.46 \cdot 10^{-10}$), and "... a scalar absolute error tolerance is quite unsuitable" (Shampine 1981). The differential equation possesses the invariant $y_2 - y_3 - y_4 = 0$, and it is recommended to use the relation $y_3' = y_2' - y_4'$ in the function subroutine (because of eventual cancellation of digits).

Originally the problem was posed on the interval $0 \leq x \leq 1000$, but Alexander (1997) discovered that the solutions possess interesting properties on a much longer interval. We follow this suggestion and consider output values at $x_{\text{out}} = 10, 10^3, 10^5, 10^7, \dots, 10^{13}$.

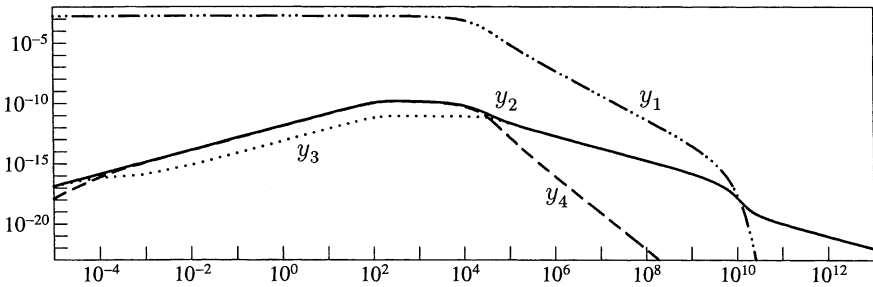


Fig. 10.1. Solution of (10.5) in double logarithmic scale

PLATE — this is a linear and non-autonomous example of medium stiffness and medium size. It describes the movement of a rectangular plate under the load of a car passing across it:

$$\frac{\partial^2 u}{\partial t^2} + \omega \frac{\partial u}{\partial t} + \sigma \Delta \Delta u = f(x, y, t). \quad (10.6)$$

The plate $\Omega = \{(x, y) ; 0 \leq x \leq 2, 0 \leq y \leq 4/3\}$ is discretized on a grid of 8×5 interior points $x_i = ih, y_j = jh, h = 2/9$ with initial and boundary conditions

$$u|_{\partial\Omega} = 0, \quad \Delta u|_{\partial\Omega} = 0, \quad u(x, y, 0) = 0, \quad \frac{\partial u}{\partial t}(x, y, 0) = 0. \quad (10.7)$$

The integration interval is $0 \leq t \leq 7$. The load $f(x, y, t)$ is idealized by the sum of two Gaussian curves which move in the x -direction and which reside on “four wheels”

$$f(x, y, t) = \begin{cases} 200(e^{-5(t-x-2)^2} + e^{-5(t-x-5)^2}) & \text{if } y = y_2 \text{ or } y_4 \\ 0 & \text{for all other } y. \end{cases}$$

The plate operator $\Delta \Delta$ is discretized via the standard “computational molecule”

$$\begin{array}{ccccc} & & 1 & & \\ & 2 & -8 & 2 & \\ 1 & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{array}$$

and the friction and stiffness parameters are chosen as $\omega = 1000$ and $\sigma = 100$. The resulting system is then of dimension 80 with negative real as well as complex eigenvalues ranging between $-500 \leq \operatorname{Re} \lambda < 0$ with maximal angle $\alpha \approx 71^\circ$ (see Definition 3.9).

BEAM — the elastic beam (1.10) of Sect. IV.1. We choose $n = 40$ in (1.10') so that the differential system is of dimension 80, and $0 \leq t \leq 5$ as integration interval. The eigenvalues of the Jacobian are purely imaginary and vary between $-6400i$ and $+6400i$ (see Eq. (2.23)). The initial conditions (1.18) and (1.19) are chosen such that the solution nevertheless appears to be smooth. However, a detailed numerical study shows that the exact solution possesses high oscillations with period $\approx 2\pi/6400$ and amplitude $\approx 10^{-6}$ (see Fig. 10.2).

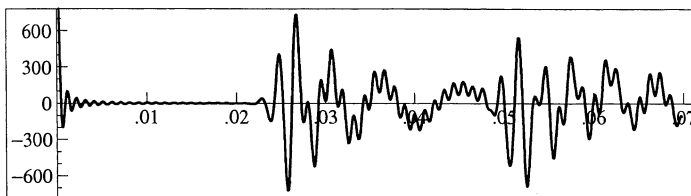


Fig. 10.2. Third finite differences $\Delta^3 y_{80} / \Delta x^3$ of solutions of the beam equation (1.10') with $n = 40$ for $0 \leq x \leq 0.07$

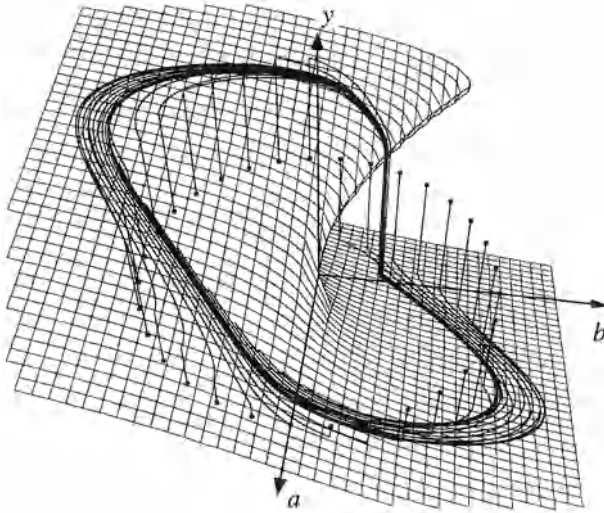


Fig. 10.3. The cusp catastrophe with $N = 32$.

CUSP — this is a combination of Zeeman's "cusp catastrophe" model ($-\varepsilon \dot{y} = y^3 + ay + b$) for the nerve impulse mechanism (Zeeman 1972) combined with the van der Pol oscillator (see Fig. 10.3)

$$\begin{aligned}\frac{\partial y}{\partial t} &= -\frac{1}{\varepsilon}(y^3 + ay + b) + \sigma \frac{\partial^2 y}{\partial x^2} \\ \frac{\partial a}{\partial t} &= b + 0.07v + \sigma \frac{\partial^2 a}{\partial x^2} \\ \frac{\partial b}{\partial t} &= (1 - a^2)b - a - 0.4y + 0.035v + \sigma \frac{\partial^2 b}{\partial x^2}\end{aligned}\tag{10.8}$$

where

$$v = \frac{u}{u + 0.1}, \quad u = (y - 0.7)(y - 1.3).$$

We put $\sigma = 1/144$ and make the problem stiff by choosing $\varepsilon = 10^{-4}$. We discretize the diffusion terms by the method of lines

$$\begin{aligned}\dot{y}_i &= -10^4(y_i^3 + a_i y_i + b_i) + D(y_{i-1} - 2y_i + y_{i+1}) \\ \dot{a}_i &= b_i + 0.07v_i + D(a_{i-1} - 2a_i + a_{i+1}) \\ \dot{b}_i &= (1 - a_i^2)b_i - a_i - 0.4y_i + 0.035v_i + D(b_{i-1} - 2b_i + b_{i+1})\end{aligned}\quad i = 1, \dots, N\tag{10.8'}$$

where

$$N = 32, \quad v_i = \frac{u_i}{u_i + 0.1}, \quad u_i = (y_i - 0.7)(y_i - 1.3), \quad D = \sigma N^2 = \frac{N^2}{144},$$

with periodic boundary conditions

$$\begin{aligned}y_0 &:= y_N, & a_0 &:= a_N, & b_0 &:= b_N, \\ y_{N+1} &:= y_1, & a_{N+1} &:= a_1, & b_{N+1} &:= b_1,\end{aligned}$$

and obtain a system of dimension $3 \cdot N = 96$. We take the initial values

$$y_i(0) = 0, \quad a_i(0) = -2 \cos\left(\frac{2i\pi}{N}\right), \quad b_i(0) = 2 \sin\left(\frac{2i\pi}{N}\right) \quad i = 1, \dots, N.$$

and $t_{\text{out}} = 1.1$.

BRUSS — this is the equation (1.6') with $\alpha = 1/50$, the same initial conditions as in Sect. IV.1, and integration interval $0 \leq t \leq 10$. But we now let $N = 500$ so that (1.6') becomes a system of 1000 differential equations with largest eigenvalue close to -20000 . The equations therefore become considerably stiff. The Jacobian of this system is banded with upper and lower bandwidth 2 (if the solution components are ordered as $u_1, v_1, u_2, v_2, u_3, v_3$, etc.).

KS — is the one-dimensional Kuramoto-Sivashinsky equation

$$\frac{\partial U}{\partial t} = -\frac{\partial^2 U}{\partial x^2} - \frac{\partial^4 U}{\partial x^4} - \frac{1}{2} \frac{\partial U^2}{\partial x} \quad (10.9)$$

with periodic boundary conditions $u(x+L, t) = u(x, t)$, taken from Collet, Eckmann, Epstein & Stubbe (1993). We choose $L = 2\pi/q$, $q = 0.025$, and take as initial condition

$$U(x, 0) = 16 \cdot \max(0, \eta_1, \eta_2, \eta_3, \eta_4),$$

$$\begin{aligned} \eta_1 &= \min(x/L, 0.1 - x/L), \\ \eta_2 &= 20(x/L - 0.2)(0.3 - x/L), \\ \eta_3 &= \min(x/L - 0.6, 0.7 - x/L), \\ \eta_4 &= \min(x/L - 0.9, 1 - x/L), \end{aligned}$$

The inverse heat equation term $-\partial^2 U / \partial x^2$ creates instability, which is stabilized for the higher oscillations by the beam equation term $-\partial^4 U / \partial x^4$. The nonlinear transport term $\partial U^2 / \partial x$ couples the modes and ensures that the solution remains bounded. All this creates wonderful chaos (see Fig. 10.4).

We solve Eq. (10.9) using the pseudo-spectral method, i.e., we consider the Fourier coefficients

$$\widehat{U}_j(t) = \frac{1}{L} \int_0^L U(x, t) e^{-iqjx} dx, \quad U(x, t) = \sum_{j \in \mathbb{Z}} \widehat{U}_j(t) e^{iqjx}, \quad (10.10)$$

so that (10.9) takes the form of an infinite dimensional ordinary differential equation

$$\widehat{U}'_j = ((qj)^2 - (qj)^4) \widehat{U}_j - \frac{iqj}{2} (\widehat{U \cdot U})_j.$$

We truncate this system as follows: for a fixed N , say $N = 1024$, we consider the N -periodic sequence $u(t) = \{u_j(t)\}$ solving the ordinary differential equation

$$u' = (d^2 - d^4)u - \frac{id}{2} \mathcal{F}_N(\mathcal{F}_N^{-1}u \cdot \mathcal{F}_N^{-1}u), \quad (10.11)$$

where d denotes the N -periodic sequence given by $d_j = qj$ for $|j| < N/2$ and $d_{N/2} = 0$, and the product of sequences in (10.11) is componentwise. The discrete

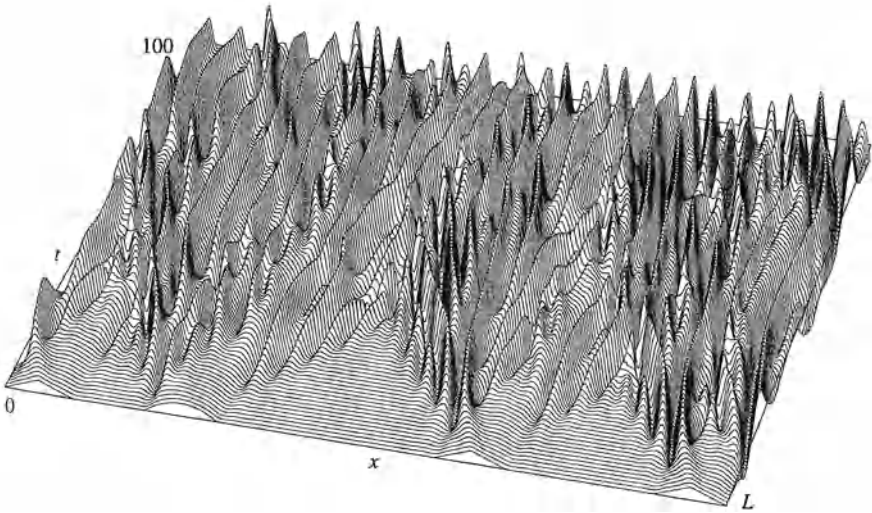


Fig. 10.4. Solution of Kuramoto-Sivashinsky equation

Fourier transform \mathcal{F}_N can be computed by FFT. From the fact that $U(x, t)$ is real it follows that the sequence u is hermitian, i.e., $u_{-j} = \bar{u}_j$. Hence, the routine REALFT from Press, Flannery, Teukolsky & Vetterling (1986, 1989), Chapter 12, is best suited for computing the right-hand side of (10.11). Since $d_0 = d_{N/2} = 0$, the components $u_0(t)$ and $u_{N/2}(t)$ are constant and need not be integrated. We thus are concerned with an ordinary differential equation of real dimension $N - 2 = 1022$. As initial values we take the discrete Fourier transform of $\{U(jL/N, 0)\}$ with the $(N/2)$ th component put to zero. In our tests we solve the differential equation (10.11) on the interval $0 \leq t \leq 100$ (see Fig. 10.4).

It can be seen from Fig. 10.5 that the Fourier modes tend to zero for $j \rightarrow \infty$, behave chaotically, and, by computing their mean values over a long period, that the modes for $qj \approx \sqrt{2}/2$ are dominant.

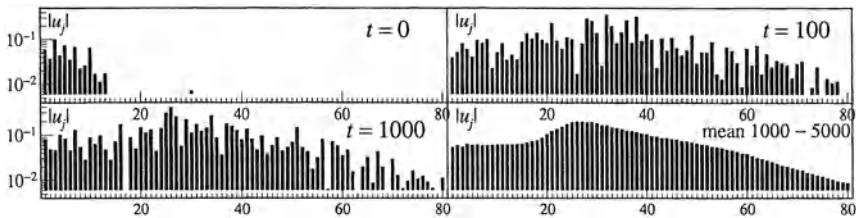


Fig. 10.5. Fourier modes for Kuramoto-Sivashinsky equation

BECKDO — the Becker-Döring model describes the dynamics of a system with a large number of identical particles which can coagulate to form clusters. We let y_k denote the expected number of k -particle clusters per unit volume. Assuming that

clusters can gain or loose only single particles, we are led to the system

$$\begin{aligned} y_1' &= -J_1 - \sum_{k=1}^{N-1} J_k, & y_N' &= J_{N-1} \\ y_k' &= J_{k-1} - J_k, & k &= 2, 3, \dots, N-1, \end{aligned} \quad (10.12)$$

where $J_k = y_1 y_k - b_{k+1} y_{k+1}$ and $b_{k+1} = \exp(k^{2/3} - (k-1)^{2/3})$. For a detailed description of this system we refer to the article by Carr, Duncan & Walshaw (1995). This equation is especially interesting because of its *metastability* (extremely slow variations in the solution over very long time intervals; see Fig. 10.6).

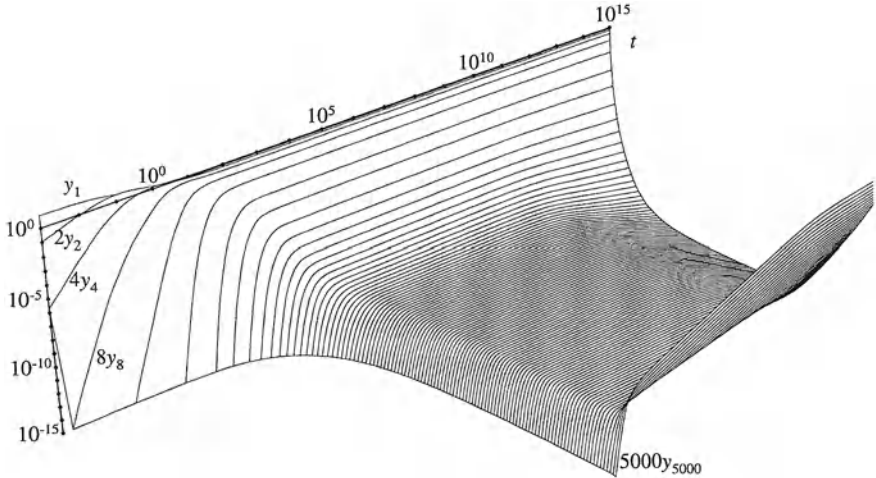


Fig. 10.6. Solutions of Becker-Döring equation (10.12)

As initial condition we take

$$y_1(0) = \varrho, \quad y_k(0) = 0 \quad \text{for } k = 2, \dots, N \quad (10.13)$$

(no clusters at the beginning). It can be seen by differentiation that the density (total number of particles per unit volume)

$$\sum_{k=1}^N k y_k \quad (= \varrho) \quad (10.14)$$

is an invariant of the system (10.12). Most numerical schemes (in particular Runge-Kutta methods and multistep methods) preserve automatically such linear invariants in the absence of round-off errors. Whenever the relation (10.14) is not satisfactorily preserved, there is the possibility to re-establish it during the computations by projections (see “differential equations with invariants”, Sect. VII.2). This precautionary measure was not used in the subsequent numerical tests.

In order to be able to observe the metastable states of the system, the dimension N has to be sufficiently large. Following the experiments of Carr, Duncan &

Walshaw (1995) we take $N = 5000$ and $\varrho = 7.5$, and consider the solution on the interval $0 \leq t \leq 10^{15}$. We compare the errors at $x_{\text{out}} = 1, 10, 10^2, 10^3, \dots, 10^{15}$.

The Jacobian of this system is tri-diagonal with an additional non-zero first row and a non-zero first column. A Gershgorin test reveals that its eigenvalues can not go, except for the initial phase, beyond -10 . Stiffness, in this example, is therefore not created by large eigenvalues of J , but by the extremely long integration interval.

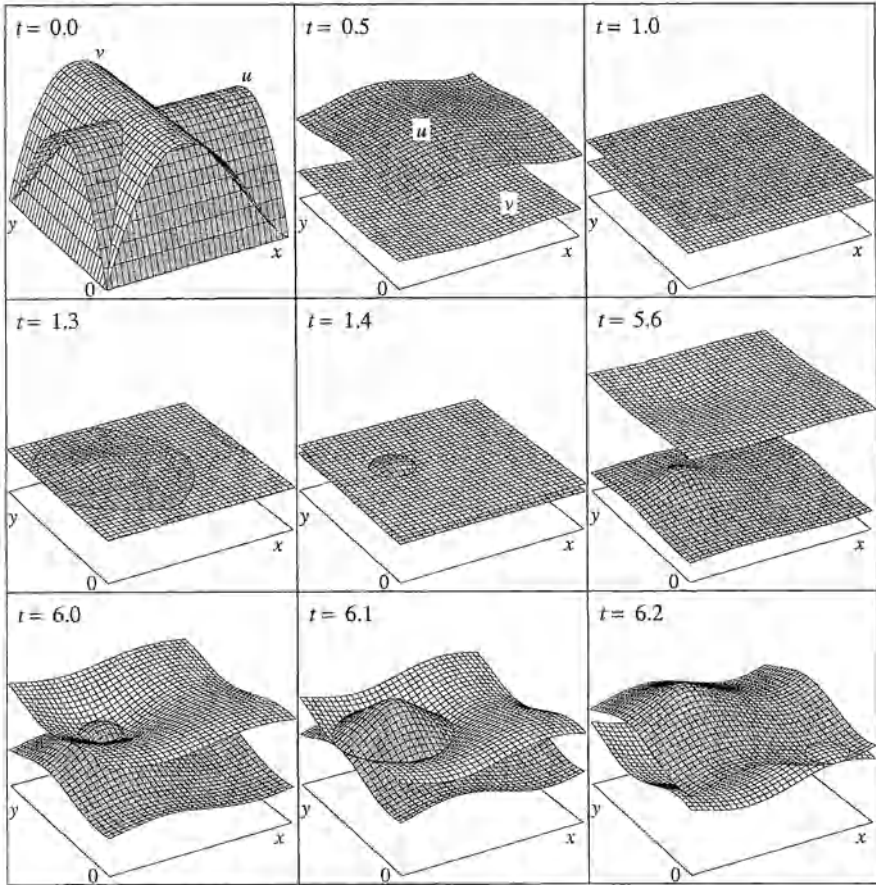


Fig. 10.7. Solution of Brusselator in 2 dimensions

BRUSS-2D — the two-dimensional Brusselator reaction-diffusion problem of Sect. II.10

$$\begin{aligned}\frac{\partial u}{\partial t} &= 1 + u^2 v - 4.4u + \alpha \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + f(x, y, t) \\ \frac{\partial v}{\partial t} &= 3.4u - u^2 v + \alpha \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right)\end{aligned}\tag{10.15}$$

in its discretized form (II.10.14), but this time we make the problem stiff by increasing the coefficient α (which was 0.002) to $\alpha = 0.1$ and by increasing the number of grid points to $N = 128$. This gives an ordinary differential equation of dimension $2N^2 = 32768$. The initial conditions, chosen here as

$$u(x, y, 0) = 22 \cdot y(1 - y)^{3/2}, \quad v(x, y, 0) = 27 \cdot x(1 - x)^{3/2}, \quad (10.16)$$

are quickly wiped out by the strong diffusion (see Fig. 10.7 for $t = 1$), we therefore suppose that the inhomogeneity $f(x, y, t)$ defined by

$$f(x, y, t) = \begin{cases} 5 & \text{if } (x - 0.3)^2 + (y - 0.6)^2 \leq 0.1^2 \text{ and } t \geq 1.1 \\ 0 & \text{else} \end{cases}$$

models an extra addition of substance u in a small disc. In order to be able to solve the linear algebra comfortably by a double FFT routine we replace the Neumann conditions of Sect. II.10 by periodic boundary conditions

$$u(x + 1, y, t) = u(x, y, t), \quad u(x, y + 1, t) = u(x, y, t).$$

As output points we choose $x_{\text{out}} = 1.5$ and 11.5 .

Results and Discussion

For each of these examples we have computed very carefully the exact solution at the specified output points. Then, the above codes have been applied with many different tolerances

$$\text{Tot} = 10^{-2-m/4}, \quad m = 0, 1, 2, \dots, 32.$$

More precisely, we set the relative error tolerance to be $\text{Rtol} = \text{Tot}$ and the absolute error tolerance $\text{Atol} = 10^{-6} \cdot \text{Tot}$ for the problems OREGO and ROBER, $\text{Atol} = 10^{-4} \cdot \text{Tot}$ for HIRES, $\text{Atol} = 10^{-3} \cdot \text{Tot}$ for PLATE and BECKDO, $\text{Atol} = 1.7 \cdot 10^{-24}$ for E5, and $\text{Atol} = \text{Tot}$ for all other problems. Several codes returned numerical results which were considerably less precise than the required precision, while other methods turned out to be more reliable. As a reasonable measure of efficiency we have therefore chosen to compare

- the actual *error* (a norm taken over all components and all output points)
- the *computing time* (of a SUN Sparc 20 Workstation) in seconds.

The obtained data are then displayed as a polygonal line in a “precision-work diagram” in double logarithmic scales. The integer-exponent tolerances 10^{-2} , 10^{-3} , 10^{-4} , \dots are displayed as enlarged symbols. The symbol for $\text{Tot} = 10^{-5}$ is specially distinguished by its gray colour. The more this line is to the right, the higher was the obtained precision; the higher this line is to the top, the slower was the code. The “slope” of the curve expresses the (effective) order of the formula: lower order methods are steeper than higher order methods. The results of the above codes on the 12 test examples are displayed in Figs. 10.8 and 10.9.

VDPOL, ROBER, OREGO — are very stiff problems of small dimension. We see from Fig. 10.8 that the Rosenbrock code RODAS is best for low tolerances (10^{-2} to 10^{-5}), whereas the extrapolation code SEULEX is superior for stringent tolerances. Due to the cheapness of the function evaluations the multistep code LSODE requires in general slightly more computing time than the one-step codes. We also remark that for a given tolerance (the position of the gray symbol for $Tol = 10^{-5}$) the code RADAU5 gives the precisest result, followed by RODAS, SEULEX, and LSODE.

HIRES — this problem is less stiff and can also be solved by explicit methods. The computing times for the explicit code DOPRI5 are initially perfectly horizontal. This is, of course, no surprise, because the step size is restricted by stability. The (explicit, but stabilized) Runge-Kutta-Chebyshev code RKC shows a considerable improvement over DOPRI5 for low tolerances. The stiff codes are still more efficient.

E5 — is a stiff and badly scaled problem, which is integrated over a very long time. Codes cannot work correctly, if the absolute tolerance $Atol$ is too large. The codes RODAS (for low tolerances) and RADAU5 (for $Tol \leq 10^{-4}$) give the best results. LSODE works safely only for $Tol \leq 10^{-5}$, whereas SEULEX has problems with round-off errors at high precision.

PLATE and BEAM — are both problems of the type $y'' = f(x, y, y')$, implemented as the first order system $y' = v$, $v' = f(x, y, v)$. For stiff codes the linear systems to be solved have a matrix of the form

$$\begin{pmatrix} \alpha I & I \\ B & C \end{pmatrix} \quad (10.17)$$

(where I is the identity matrix). Using the option $IWORK(9)=N/2$ (where N is the dimension of the first order system) our codes do the first $N/2$ elimination sweeps analytically and the dimension of the linear system is halved. Without this option, the computing times for the codes RADAU5, RODAS, and SEULEX would be larger by a factor of about 3.0, 1.7, and 2.6, respectively (these numbers are for the BEAM problem at $Tol = 10^{-5}$). We did not include here the results of LSODE, for which we did not have an easy possibility for such a reduction. For the PLATE problem we also exploited the banded structure of $\partial f / \partial y$ and $\partial f / \partial v$ by putting $MLJAC=16$ and $MUJAC=16$.

For both problems the explicit code DOPRI5 was applicable too. A curious phenomenon arose for DOPRI5 at the PLATE problem: as expected, for low tolerance requirements ($Tol \geq 10^{-5}$), the code appeared to be restricted by stability, gave computing times independent of Tol and issued the message “the problem seems to be stiff”. But for more stringent tolerances the code was restricted by precision, with computing times unexpectedly high above those of the implicit code RADAU5. The analysis of Sect. IV.15 for the Prothero & Robinson problem (15.1) gives an

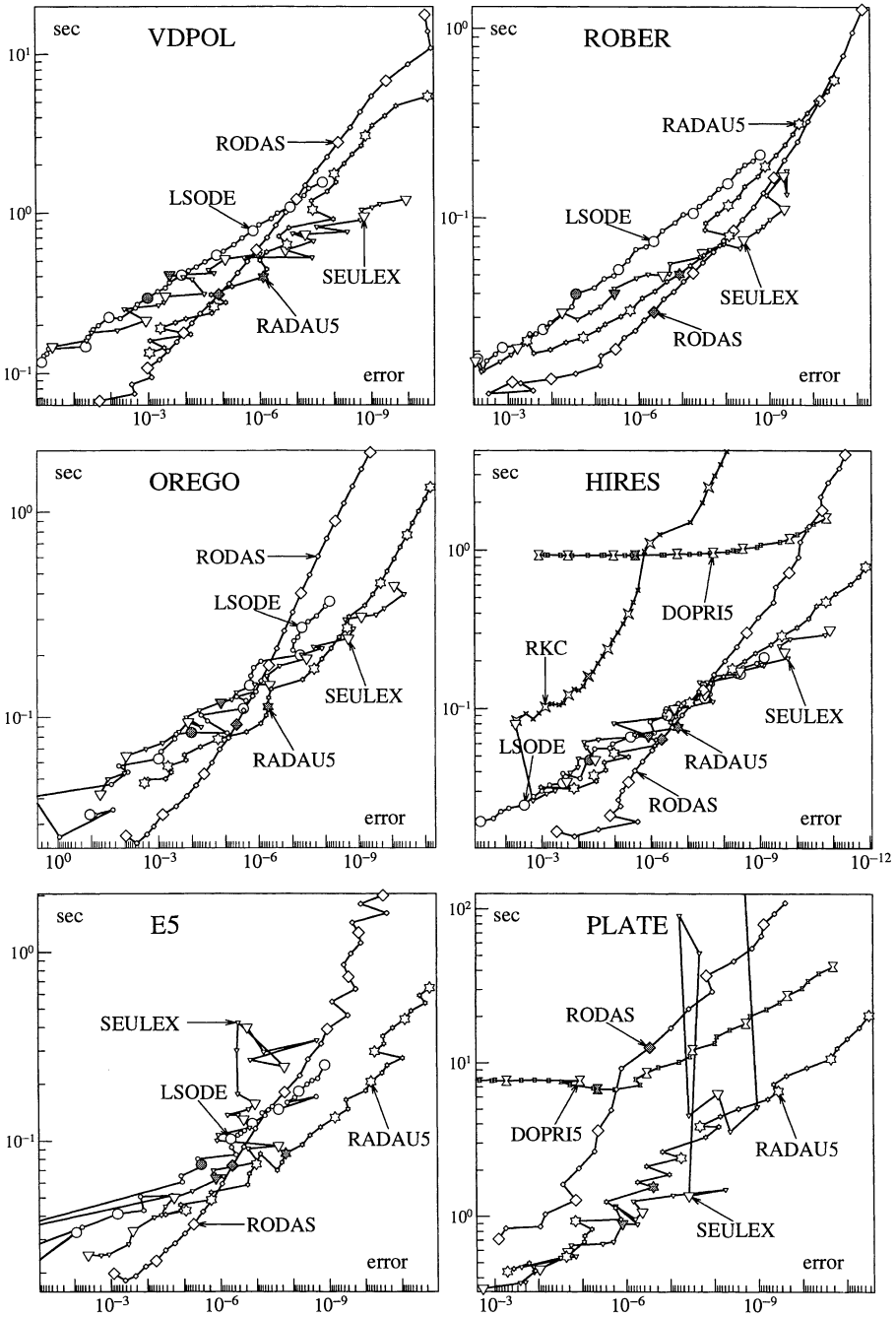


Fig. 10.8. Work-precision diagrams for problems of dimension 2 to 80

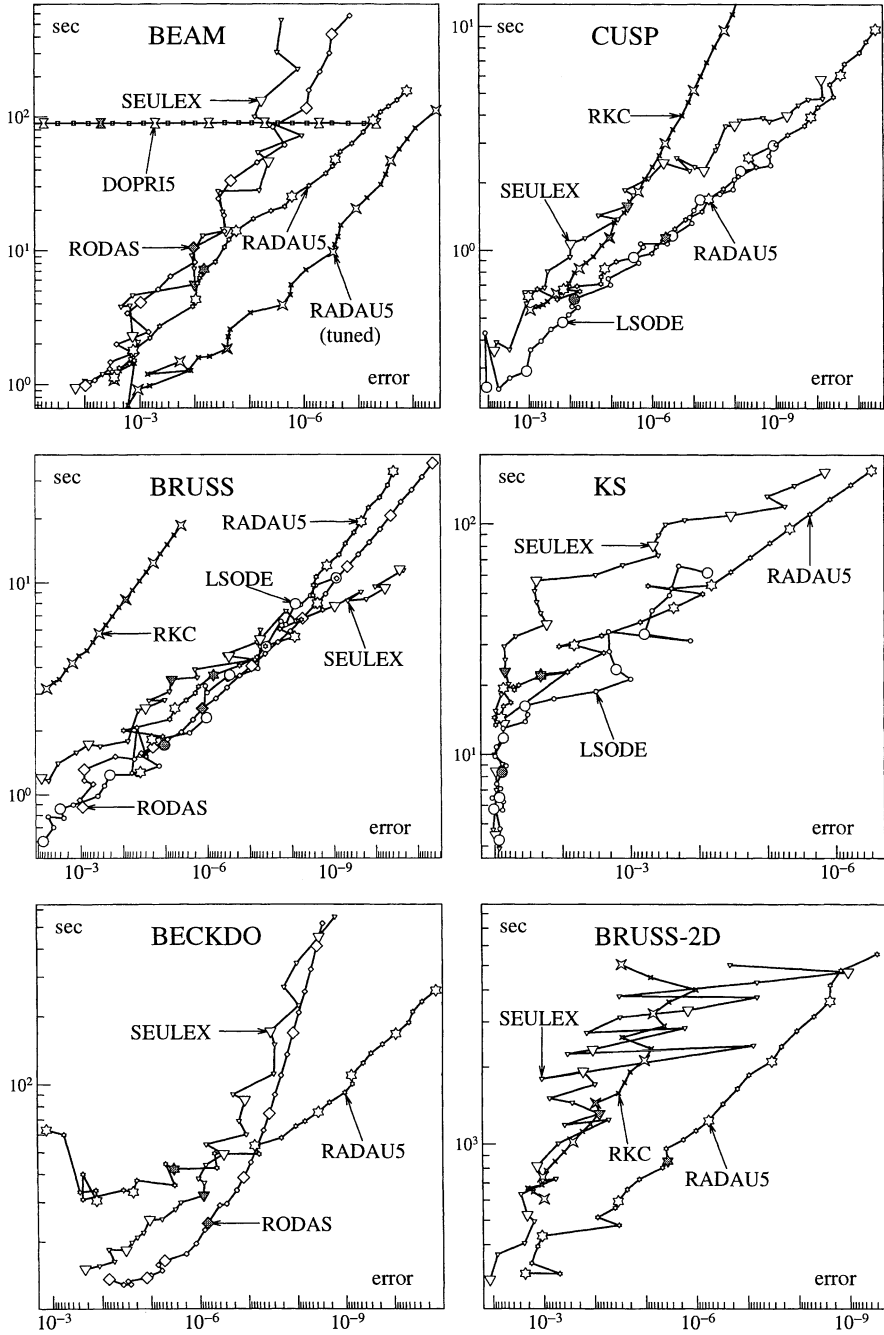


Fig. 10.9. Work-precision diagrams for problems of dimension 80 to 32768

explanation for this fact. We see that stiff problems not only create loss of stability, but also loss of precision for explicit integrators.

Especially for the BEAM problem, a problem with expensive linear algebra, the efficiency of the codes can be considerably increased by *tuning the parameters*. If, for the integration with RADAU5, we put

WORK(3)=0.1	(Jacobian less often recomputed)
WORK(4)=0.3	(Newton iterations stopped earlier)
WORK(5)=0.99	} (Step size changed less often,
WORK(6)=2.	
	decreasing number of LU-decompositions)

then the computing time decreases by a factor between 2 and 5. Fig. 10.9 shows the spectacular improvement of this “tuned” run.

CUSP — the Jacobian of this problem is of the form

$$J = \begin{pmatrix} A_1 & B_1 & & D_1 \\ C_2 & A_2 & \ddots & \\ & \ddots & \ddots & B_{N-1} \\ D_N & & C_N & A_N \end{pmatrix} \quad (10.18)$$

where A_i, B_i, C_i, D_i are 3×3 matrices, and an efficient solution of the linear system needs a special treatment (see Exercise 1). However the considered methods, with the exception of the Rosenbrock methods, do not require an exact Jacobian. Therefore, an easy possibility for a considerable reduction of computing time is simply to use the codes in the banded version by putting $ML=MU=3$. The D_1 and D_N are neglected and we obtain the computing times displayed in Fig. 10.9. If the Jacobian were treated as a full matrix, the computing times would increase by a factor of 8.3, 6.6, and 4.8 for the codes RADAU5, SEULEX, and LSODE, respectively (these numbers are for $Tol = 10^{-5}$). The explicit code RKC gives excellent results for low precision, whereas the results of DOPRI5 (more than 30 seconds) are outside of the picture.

BRUSS — for this one-dimensional reaction-diffusion problem the linear algebra is done in the “banded” version with “analytical Jacobian”. The problem is very stiff (large diffusion constant and small Δx) and an explicit method, such as DOPRI5, would require close to 60000 steps of integration. The code RKC works well, although less efficiently than the stiff integrators.

KS — the solution of this problem is sensitive with respect to changes in the initial values, a phenomenon already encountered in the LRNZ problem of Sect. II.10. Similarly as there, the precision increases only for Tol beyond a certain threshold. The Jacobian of this problem is full. Numerical experiments revealed that the codes worked best when the Jacobian is replaced by a diagonal matrix with $(qj)^2 - (qj)^4$ in its j th entry. Rosenbrock methods, which require an exact Jacobian, are not efficient here. The explicit codes RKC and DOPRI5 need too much computing time.

BECKDO — for this problem, the stiff codes (the only ones which work) require the solution of linear systems of the form

$$\begin{pmatrix} u & v^T \\ w & T \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}, \quad (10.19)$$

where v , w , b are $(n-1)$ -dimensional vectors and T is a tri-diagonal matrix. Since the linear algebra routines are completely separated from the codes RADAU5, RODAS and SEULEX, it is easy to replace these routines by a special program which solves (10.19) efficiently as follows

$$\begin{aligned} x &= (a - v^T T^{-1} b) / (u - v^T T^{-1} w) \\ y &= T^{-1} b - x T^{-1} w. \end{aligned} \quad (10.20)$$

It is not necessary to alter the stiff integrator itself.

Fig. 10.9 shows that, as usual, RODAS is best for low tolerances and RADAU5 is preferable for high precision. *Not* as usual is the fact that RODAS performs very badly for stringent tolerances. We explain this by the fact that the linear system (10.19) is sensitive to round-off errors, or, as Wilkinson would turn it, delivers a solution for a *wrong* Jacobian. Thus, the order of the Rosenbrock method drops to 1.

BRUSS-2D — due to its large dimension ($n = 2 \cdot 128^2 = 32768$), this problem makes no sense in full or even banded linear algebra. We therefore solved the linear equations (in the codes with separated linear algebra, see the corresponding remarks in the BECKDO problem) by FFT methods, taking into account only the (stiff) diffusion terms and neglecting the (in this problem non-stiff) reaction terms. The FFT codes used were those of Press, Flannery, Teukolsky & Vetterling (1986, 1989) in the chapter on partial differential equations. A special advantage of the Radau method is here that the complex algebra, which is anyway used in FFT, crunches the complex eigenvalues of the Runge-Kutta matrix without further harm.

For this problem, which is a typical parabolic partial differential equation with non-stiff nonlinearities, we have made a detailed comparison of the performances of the implicit code RADAU5, the “stabilized” explicit code RKC, and the explicit code DOPRI5, in dependence of the discretization parameter $\Delta x = \Delta y = 1/N$ and the diffusion parameter α (see Eqs. (10.15) and (II.10.14)). The results (number of function calls and computing times) are displayed in Table 10.1, where the best performances are displayed in boldface characters. We can see how the olympic fire goes over from DOPRI5, which is best for low stiffness ($\alpha N^2 \leq 1$), by increasing the stiffness first to RKC, and then (for $\alpha N^2 \geq 1000$) to the implicit RADAU5 code. We also observe that the number of function evaluations is nearly independent of the stiffness for RADAU5, behaves like $Const \cdot \sqrt{\alpha} \cdot N$ for RKC, and like $Const \cdot \alpha \cdot N^2$ for DOPRI5.

Comparisons Between Codes of the Same Type. Figs. 10.8 and 10.9, which are a sort of “Final Competition of Wimbledon”, contain only one code from each class of integration methods (Radau methods, Implicit Runge-Kutta, Rosenbrock,

Table 10.1. Function evaluations / computing times at $Tol = 10^{-5}$

RADAU5	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
$\alpha = 10^{-3}$	3372/19.8	3233/84.9	3271/413.5	3290/2215.6	3261/14902.1
$\alpha = 10^{-2}$	1286/7.7	1322/36.2	1295/167.4	1381/868.8	1380/6459.3
$\alpha = 10^{-1}$	1150/6.8	1131/30.9	1227/ 172.3	1173/ 854.9	1204/ 5664.9
$\alpha = 1$	1195/7.8	1199/ 33.0	1247/ 177.3	1242/ 945.9	1258/ 5961.2
RKC	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
$\alpha = 10^{-3}$	2367/4.7	2277/18.6	2249/76.3	2311/ 352.5	2911/ 1912.0
$\alpha = 10^{-2}$	1661/3.2	1674/ 13.8	2078/ 70.4	3379/ 511.5	6259/ 4086.9
$\alpha = 10^{-1}$	1899/ 3.6	2823/ 22.5	5047/ 176.8	9666/1446.2	18911/12312.2
$\alpha = 1$	4013/ 7.2	7565/58.9	14631/503.4	29022/4328.8	
DOPRI5	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
$\alpha = 10^{-3}$	976/ 2.0	1030/ 8.5	1408/ 48.5	3286/509.4	11464/7704.2
$\alpha = 10^{-2}$	784/ 1.6	1894/15.4	6976/240.6	27478/4369.6	
$\alpha = 10^{-1}$	4366/9.0	17176/145.5	68446/2419.7	273568/43982.2	
$\alpha = 1$	42832/90.6	171010/1505.8	683836/24362.7		

and extrapolation methods). Following are some comparisons within each of these classes.

Radau Methods. For a comparison of Radau methods of various orders (see also the results of Reymond (1989) in the first edition), we have written a code RADAUP, which allows to choose with the help of a method flag `IWORK(11)=3,5,7` to choose between $s = 3, 5$, or 7 (i.e., between orders $p = 5, 9$, or 13). The code is for $s = 3$ mathematically equivalent to RADAU5, but, due to a different coding, slightly slower. We can see in Fig. 10.10 how the higher order pays off for higher precision, but for lower precision arise problems due to large step sizes and bad convergence of the Newton iterations.

Implicit Runge-Kutta Methods. It has for a long time been taken for granted that only DIRK and SDIRK methods could be implemented efficiently. Our experience shows that the diagonally implicit method SDIRK4, constructed in Section IV.6, gives rather disappointing results (see Fig. 10.11). An exception is the BEAM problem with its, microscopically, highly oscillatory solutions. Since the code SDIRK4 has not the option for “second order” linear algebra, we have also applied RADAU5 without this option. The computing times for RADAU5 are therefore not the same as in Fig. 10.9.

Rosenbrock Methods. There is usually not much difference between the performance of the different Rosenbrock methods (see Fig. 10.12). In spite of their larger number of stages, the codes RODAS5 (order 5) and RODAS (order 4) give often

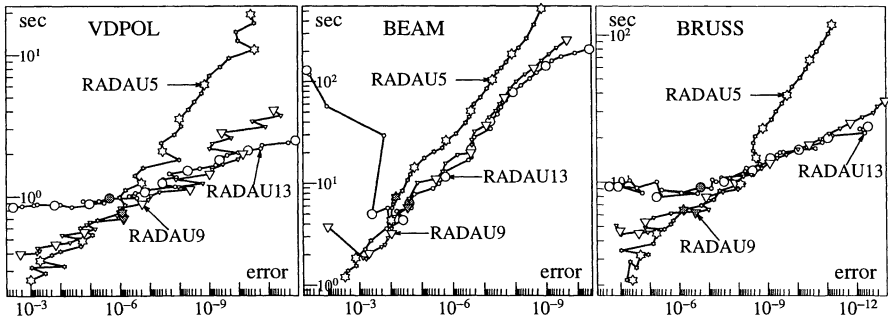


Fig. 10.10. Comparison between Radau codes

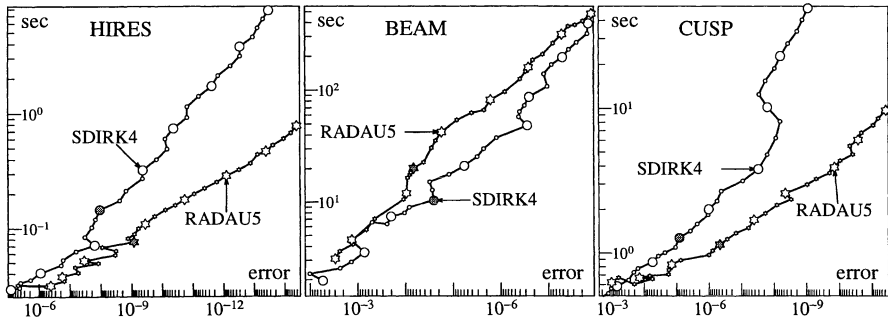


Fig. 10.11. Comparison between implicit Runge-Kutta codes

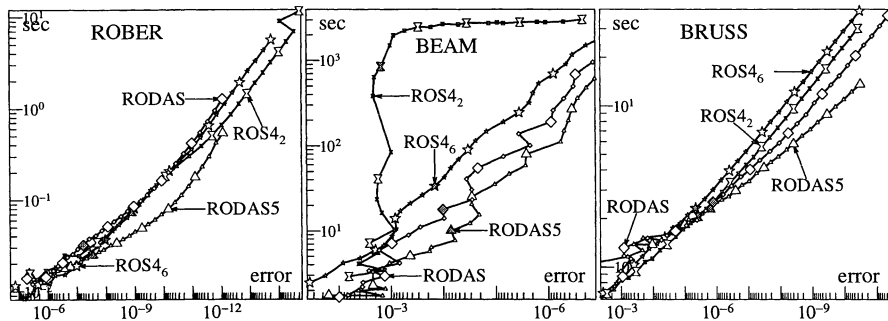


Fig. 10.12. Comparison between Rosenbrock codes

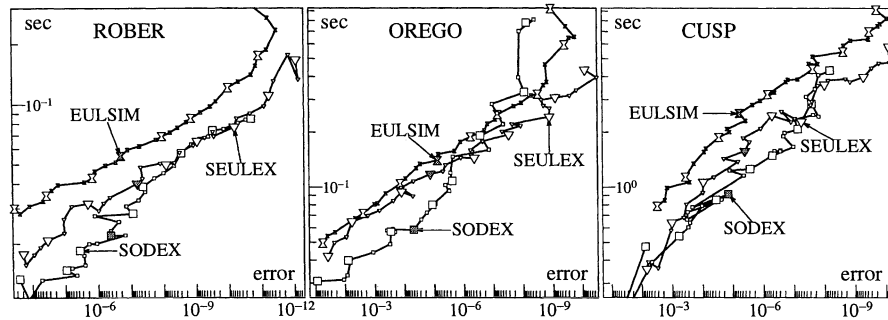


Fig. 10.13. Comparison between extrapolation codes

the best results. Among the 4th order “classical” Rosenbrok methods of Table 7.2 the best is in general “method 2” with its small error constant; it fails completely, however, on the Beam problem due to lack of A -stability. “Method 6” corresponds to the choice of coefficients which give an L -stable method.

Extrapolation Methods. The code SODEX, which is based on an h^2 -extrapolation of the semi-implicit midpoint rule, is clearly superior to SEULEX for low precision (see Fig. 10.13). The opposite situation appears for more stringent tolerances; here we observe an order reduction phenomenon, which is explained in Sect. VI.5 below. We have also included in these tests the results of the code EULSIM by Deuflhard, Novak & Poehle (poehle@sc.zib-berlin.de) which is another implementation of the extrapolated semi-implicit Euler method, with a different stepsize sequence.

Chebyshev Methods. During the final realization of these experiments we have received a code DUMKA3 (written by A. Medovikov, nuirect@inm.ras.ru) which implements an extension of the optimal Chebyshev methods of Lebedev (see Sect. IV.2) to third order. This code is still in a very experimental stage, but the results, presented in Fig. 10.14, are very promising.

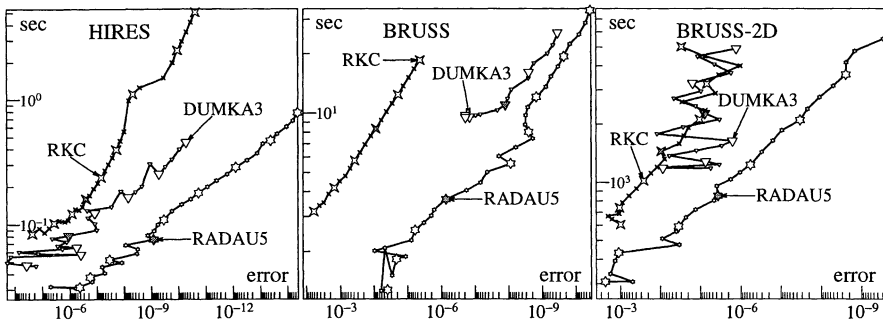


Fig. 10.14. Comparison between Chebyshev codes and RADAU5

Partitioning and Projection Methods

Most codes for solving stiff systems . . . spend most of their time solving systems of linear equations . . .

(Watkins & HansonSmith 1983)

Further spectacular reductions of the work for the linear algebra are often possible. One of the oldest ideas is to *partition* a stiff system into a (hopefully) small stiff system and a large nonstiff part,

$$\begin{aligned} y'_a &= f_a(y_a, y_b) & (\text{stiff}) \\ y'_b &= f_b(y_a, y_b) & (\text{nonstiff}), \end{aligned} \tag{10.21}$$

so that the two systems can be treated by two different methods, one implicit and the other explicit (e.g. Hofer 1976). The theory of P -series in Sect. II.14 had its

origin in the study of the order properties of such methods. A difficulty of this approach is, of course, to decide *which* equations should be the stiff ones. Further, stiffness may affect subspaces which are *not* parallel to the coordinate axes. We shall therefore turn our attention to procedures which do not adapt the underlying *numerical method* to the partitioning, but the *linear algebra* only. An excellent survey of the older literature on these methods is given by Söderlind (1981). The following definition describes an especially promising class of problems:

Definition 10.1 (Björck 1983, 1984). The system $y' = f(x, y)$ is called *separably stiff* at a position x_0, y_0 if the Jacobian $J = \frac{\partial f}{\partial y}(x_0, y_0)$ possesses $k < n$ eigenvalues $\lambda_1, \dots, \lambda_k$ such that

$$\min_{1 \leq i \leq k} |\lambda_i| \gg \max_{k+1 \leq i \leq n} |\lambda_i|.$$

The eigenvalues $\lambda_1, \dots, \lambda_k$ are called the *stiff eigenvalues* and

$$\mu = \min_{1 \leq i \leq k} |\lambda_i| / \max_{k+1 \leq i \leq n} |\lambda_i| \quad (10.22)$$

the *relative separation*. The space D spanned by the *stiff eigenvectors* is called the *dominant invariant subspace*.

For example, the Robertson problem (10.2) possesses only *one* stiff eigenvalue (close to -2000), and is therefore separably stiff with $k = 1$. The CUSP problem (10.8') of dimension 96 has 32 large eigenvalues which range, except for transient phases, between -20000 and -60000 . All other eigenvalues satisfy approximately $|\lambda| < 30$. This problem is, in fact, a singular perturbation problem (see Sect. VI.1), and such problems are all separably stiff. The other large problems of this section have eigenvalues scattered all around. A.R. Curtis' study (1983) points out that in *practical* problems separably stiff problems are rather seldom.

The Method of Gear and Saad. Implicit methods such as (transformed) Runge-Kutta or multistep formulas require the solution of a linear system (where we denote, as usual in linear algebra, the unknown vector by x)

$$Ax = b \quad \text{where} \quad A = \frac{1}{h\gamma} I - J \quad (10.23)$$

with *residual* $r = b - Ax$. We choose k (usually) orthogonal vectors q_1, \dots, q_k in such a way that the span $\{q_1, \dots, q_k\} = \tilde{D}$ is an *approximation* to the dominant subspace D , and denote by Q the $k \times n$ -matrix formed by the columns q_j ,

$$Q = (q_1, \dots, q_k). \quad (10.24)$$

There are now several possibilities for replacing the solution x of (10.23) by an approximate solution $\tilde{x} \in \tilde{D}$. One of the most natural is to require (Saad 1981, Gear & Saad 1983; in fact, Galerkin 1915) that the residual of \tilde{x} ,

$$\tilde{r} = b - A\tilde{x} = A(x - \tilde{x}), \quad (10.25)$$

be *orthogonal* to \tilde{D} , i.e., that $Q^T(b - A\tilde{x}) = 0$. If we write \tilde{x} in the basis of (10.24) as $\tilde{x} = Q\tilde{y}$, this yields

$$H\tilde{y} = Q^T b, \quad (10.26)$$

where

$$H = Q^T A Q \quad \text{or} \quad QH = A Q, \quad (10.27)$$

which means that we have to solve a linear system of dimension k with matrix H . A particularly good choice for \tilde{D} is a *Krylov subspace* spanned by an arbitrary vector r_0 (usually the residual of a well chosen initial approximation x_0),

$$\tilde{D} = \text{span} \{r_0, Ar_0, A^2 r_0, \dots, A^{k-1} r_0\}. \quad (10.28)$$

The vectors (10.28) constitute the sequence created by the well-known power method. Therefore, in the case of a separably stiff system, as analyzed by D.J. Higham (1989), the space \tilde{D} approaches the space D extremely well as soon as its dimension is sufficiently high. In the *Arnoldi process* (Arnoldi 1951) the vectors of (10.28) are successively orthonormalized (Gram-Schmidt) as

$$\begin{aligned} q_1 &= r_0 / \|r_0\| \\ \hat{q}_2 &= Aq_1 - h_{11}q_1, \quad q_2 = \hat{q}_2 / h_{21} \quad \text{with} \quad h_{21} = \|\hat{q}_2\| \end{aligned}$$

and so on, and we see that

$$\begin{aligned} Aq_1 &= h_{21}q_2 + h_{11}q_1 \\ Aq_2 &= h_{32}q_3 + h_{22}q_2 + h_{12}q_1 \\ &\dots \end{aligned} \quad (10.29)$$

which, compared to (10.28), shows that H is *Hessenberg*. For A symmetric, H is also symmetric, hence tridiagonal, so that the method is equivalent to the conjugate gradient method.

Two features are important for this method: Firstly, the matrix A need never be computed nor stored. All that is needed are the matrix-vector multiplications in (10.29), which can be obtained from the “directional derivative”

$$Jv \approx [f(x, y + \delta v) - f(x, y)] / \delta. \quad (10.30)$$

Several people therefore call such methods “matrix-free”. Secondly, the dimension k does not have to be known: one simply computes one column of H after the other and periodically estimates the residual. As soon as this estimate is small enough (or k becomes too large) the algorithm stops. We also mention two variants of the method:

1. (Gear & Saad 1983, p. 595). Before starting the computation of the Krylov subspace, perform some initial iteration of the power method on the initial vector r_0 , using either the matrix A or the matrix J . Lopez & Trigiante (1989) report excellent numerical results for this procedure.

2. *Incomplete Orthogonalization* (Saad 1982). The new vector Aq_j is only orthogonalized against the previous p vectors, where p is some small integer. This

makes H a banded matrix and saves computing time and memory. For symmetric matrices, the ideal choice is of course $p = 2$, for matrices more and more unsymmetric p usually is increased to 10 or 15.

The EKBWH-Method (this tongue-twister stands for Enright, Kamel, Björck, Watkins and HansonSmith). Here, the matrices A (and J) in (10.23) are replaced by approximations

$$\tilde{A} = \frac{1}{h\gamma} I - \tilde{J} \quad (10.31)$$

where \tilde{J} should approach J sufficiently well and the matrix \tilde{A} be relatively easy to invert. \tilde{J} is determined as follows: Complete (theoretically) the vectors (10.24) to an orthogonal basis (Q, \hat{Q}) of \mathbb{R}^n . In the new basis J becomes

$$\begin{pmatrix} Q^T \\ \hat{Q}^T \end{pmatrix} J (Q, \hat{Q}) = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \quad (10.32)$$

and we have

$$Q^T J Q = T_{11}. \quad (10.33)$$

If $\text{span } Q = \tilde{D}$ approaches D , then T_{11} will contain the stiff eigenvalues and T_{21} will tend to zero. If $\tilde{D} = D$ exactly, then $T_{21} = 0$ and (10.32) is a block-Schur decomposition of J . For separably stiff systems $\|T_{22}\|$ will become small compared to $(h\gamma)^{-1}$ and we define

$$\tilde{J} = (Q, \hat{Q}) \begin{pmatrix} T_{11} & T_{12} \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Q^T \\ \hat{Q}^T \end{pmatrix} = Q(T_{11}Q^T + T_{12}\hat{Q}^T) \stackrel{(10.32)}{=} QQ^T J.$$

This shows \tilde{J} to be the orthogonal projection of J onto \tilde{D} . The inverse of \tilde{A} is computed by developing $(I - B)^{-1} = I + B + B^2 + \dots$ as a geometric series

$$\begin{aligned} \tilde{A}^{-1} &= h\gamma(I - h\gamma QQ^T J)^{-1} \\ &= h\gamma(I + h\gamma QQ^T J + h^2\gamma^2 Q \underbrace{Q^T J Q}_{T_{11}} Q^T J + \dots) \\ &= h\gamma(I + Q(h\gamma I + h^2\gamma^2 T_{11} + h^3\gamma^3 T_{11}^2 + \dots)Q^T J) \\ &= h\gamma(I + Q(\frac{1}{h\gamma} I - T_{11})^{-1} Q^T J) \end{aligned} \quad (10.34)$$

which only requires the solution of the “small” system with matrix $(I/h\gamma - T_{11})$ (the last expression is called the Sherman-Morrison-Woodbury formula).

Choice of Q :

— Björck (1983) computes the precise span of D , by Householder transforms followed by block- QR iterations. For separably stiff systems the block T_{21} converges to zero linearly with ratio μ^{-1} so that usually 2 or 3 iterations are sufficient. A disadvantage of the method is that an estimate for the dimension k of D must be known in advance.

— Enright & Kamel (1979) transform J to Hessenberg form and stop the transformations when $\|T_{21}\| + \|T_{22}\|$ become sufficiently small (remark that T_{21} is non zero in its last column only). Thus the dimension k can be discovered dynamically. Enright & Kamel combine the Householder reflexions with a pivoting strategy and repeated row & column permutations in order to make T_{22} small as fast as possible. It was first observed numerically (by Carlsson) and then shown theoretically (Söderlind 1981) that this pivoting strategy “needs some comments”: if we start from (10.32), by knowing that

$$\begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$$

is Hessenberg in its first k columns, (with $h_{21} \neq 0$, $h_{32} \neq 0, \dots$) and do the analysis of formulas (10.29) backwards, we see that the space \tilde{D} for the Enright & Kamel method is a Krylov subspace created by q_1 (D.J. Higham 1989). Thus only the first permutation influences the result.

— Watkins & HansonSmith (1983) start from an arbitrary $Q^{(0)}$ followed by several steps of the block power method

$$JQ^{(i)} = Q^{(i+1)}R^{(i+1)} \quad (10.35)$$

where $R^{(i+1)}$ re-orthogonalizes the vectors of the product $JQ^{(i)}$. A great advantage of this procedure is that no large matrix needs to be computed nor stored. The formulas (10.35) as well as (10.34) only contain matrix-vector products which are computed by (10.30). The disadvantage is that the dimension of the space must be known.

Stopping Criteria. The above methods need a criterion on the goodness of the approximation \tilde{J} to decide whether the dimension k is sufficient. Suppose that we solve the linear equation (10.23) by a modified Newton correction which uses \tilde{A} as “approximate Jacobian”

$$\tilde{x} = x_0 + \tilde{A}^{-1}(b - Ax_0),$$

then the convergence of this iteration is governed by the condition

$$\varrho(I - \tilde{A}^{-1}A) = \varrho(\tilde{A}^{-1}(\tilde{A} - A)) = \varrho(\tilde{A}^{-1}(J - \tilde{J})) < 1. \quad (10.36)$$

A reasonable condition is therefore that the spectral radius ϱ of $\tilde{A}^{-1}(J - \tilde{J})$ is plainly smaller than 1. Let us compute this value for the Björk method ($T_{21} = 0$): since the eigenvalues of a matrix C are invariant under the similarity transforma-

tion $T^{-1}CT$, we have

$$\begin{aligned}\varrho(\tilde{A}^{-1}(J - \tilde{J})) &= \varrho\left(\left(\frac{1}{h\gamma}I - \begin{pmatrix} T_{11} & T_{12} \\ 0 & 0 \end{pmatrix}\right)^{-1} \begin{pmatrix} 0 & 0 \\ 0 & T_{22} \end{pmatrix}\right) \\ &= \varrho\left(\begin{pmatrix} (\frac{1}{h\gamma}I - T_{11})^{-1} & \times \times \times \\ 0 & h\gamma I \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & T_{22} \end{pmatrix}\right) \\ &= \varrho\left(\begin{pmatrix} 0 & \times \times \times \\ 0 & h\gamma T_{22} \end{pmatrix}\right) = \varrho(h\gamma T_{22}).\end{aligned}$$

In practice, a condition of the form

$$\|h\gamma T_{22}\| < 1, \quad (10.37)$$

where $\|\cdot\|$ is usually the Frobenius norm $\sqrt{\sum_{i,j} a_{ij}^2}$, ensures a reasonable rate of convergence. For an analogous condition in the Enright-Kamel case see Exercise 3 below.

Exercises

1. (The red-black reduction). The Jacobian matrix of the (periodic) cusp catastrophe model (10.8') is of the form

$$\begin{pmatrix} A_1 & B_1 & & & C_1 \\ C_2 & A_2 & B_2 & & \\ & \ddots & \ddots & \ddots & \\ & & C_{2m-1} & A_{2m-1} & B_{2m-1} \\ B_{2m} & & & C_{2m} & A_{2m} \end{pmatrix} \quad (10.38)$$

where A_i, B_i, C_i are (3×3) -matrices. Write a solver which solves linear equations with matrix (10.38) using the “red-black ordering reduction”. This means that A_1, A_3, A_5, \dots are used as (matricial) pivots to eliminate $C_2, C_4, \dots, B_2, B_4, \dots$ above and below by Gaussian block-elimination. Then the resulting system is again of the same structure as (10.38) with *halved* dimension. If the original system's dimension contains 2^k as prime factor, this process can be iterated k times. Study the increase of performance which this algorithm allows for the RADAU5 and Rosenbrock codes on model (10.8'). The algorithm is also highly parallelizable.

2. Show by numerical experiments that the circular nerve (10.8') loses its limit cycle when the diffusion coefficient D becomes either too small (the message does not go across the water fall) or too large (the limit cycle then melts down across the origin).
3. (Stopping criterion for Enright & Kamel method; D.J. Higham 1989). Suppose that the matrix J has been transformed to partial Hessenberg form (see

(10.32))

$$\begin{pmatrix} Q^T \\ \hat{Q}^T \end{pmatrix} J(Q, \hat{Q}) = \begin{matrix} k & n-k \\ n-k \end{matrix} \begin{pmatrix} H & T_{12} \\ (0 \ b) & T_{22} \end{pmatrix}$$

where H is upper Hessenberg and b a column vector. Show that the criterion (10.36) then becomes

$$\varrho(h\gamma B) < 1$$

where

$$B = \begin{matrix} k \\ n-k \end{matrix} \begin{pmatrix} k-1 & 1+n-k \\ 0 & -h\gamma \bar{H}^{-1} T_{12}(b \ T_{22}) \\ 0 & (b \ T_{22}) \end{pmatrix}$$

with $\bar{H} = (I - h\gamma H)$. Since $\varrho(B)$ is the same as the spectral radius of its lower $1+n-k$ by $1+n-k$ principal submatrix, a sufficient condition for convergence is

$$|h\gamma| \sqrt{\|T_{22}\|^2 + \|b\|^2 + \|y\|^2} < 1$$

where y^T is the k -th row of the matrix $-h\gamma \bar{H}^{-1} T_{12}(b \ T_{22})$.

IV.11 Contractivity for Linear Problems

He who loves practice without theory is like the sailor who boards ship without a rudder and compass and never knows where he may be cast.
(Leonardo da Vinci 1452-1519,
cited from M. Kline, *Mathematical Thought* ... 1972, p. 224)

The stability analysis of the preceeding sections is based on the transformation of the Jacobian $J \approx \partial f / \partial y$ to diagonal form (see Formulas (2.5), (2.6) of Sect. IV.2). Especially for large-dimensional problems, however, the matrix which performs this transformation may be badly conditioned and destroy all the nice estimations which have been obtained.

Example 11.1. The discretization of the hyperbolic problem

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} \quad (11.1)$$

by the method of lines leads to

$$y' = Ay, \quad A = \lambda \begin{pmatrix} -1 & 1 & & \\ & -1 & \ddots & \\ & & \ddots & 1 \\ & & & -1 \end{pmatrix}, \quad \lambda = \frac{1}{\Delta x} > 0. \quad (11.2)$$

This matrix has all eigenvalues at $-\lambda$ and the above spectral stability analysis would indicate fast asymptotic convergence to zero. But neither the solution of (11.1), which just represents a travelling wave, nor the solution of (11.2), if the dimension becomes large, have this property. So our interest in this section is to obtain rigorous bounds for the numerical solution (see (2.3))

$$y_{m+1} = R(hA)y_m \quad (11.3)$$

in different norms of \mathbb{R}^n or \mathbb{C}^n . Here $R(z)$ represents the stability function of the method employed. We have from (11.3)

$$\|y_{m+1}\| \leq \|R(hA)\| \cdot \|y_m\| \quad (11.4)$$

(see Volume I, Sect. I.9, Formula (9.10)), and contractivity is assured if

$$\|R(hA)\| \leq 1.$$

Euclidean Norms (Theorem of von Neumann)

People in mathematics and science should be reminded that many of the things we take for granted today owe their birth to perhaps one of the most brilliant people of the twentieth century — John von Neumann.

(John Impagliazzo, quoted from SIAM News September 1988)

Let the considered norm be Euclidean with the corresponding scalar product denoted by $\langle \cdot, \cdot \rangle$. Then, for the solution of $y' = Ay$ we have

$$\frac{d}{dx} \|y\|^2 = \frac{d}{dx} \langle y, y \rangle = 2\operatorname{Re} \langle y, y' \rangle = 2\operatorname{Re} \langle y, Ay \rangle, \quad (11.5)$$

hence the solutions are decaying in this norm if

$$\operatorname{Re} \langle y, Ay \rangle \leq 0 \quad \text{for all } y \in \mathbb{C}^n. \quad (11.6)$$

This result is related to Theorem 10.6 of Sect. I.10, because

$$\operatorname{Re} \langle y, Ay \rangle \leq \mu_2(A) \|y\|^2, \quad (11.7)$$

where $\mu_2(A)$ is the logarithmic norm of A (Eq. (10.20) of Sect. I.10).

Theorem 11.2. *Let the rational function $R(z)$ be bounded for $\operatorname{Re} z \leq 0$ and assume that the matrix A satisfies (11.6). Then, in the matrix norm corresponding to the scalar product we have*

$$\|R(A)\| \leq \sup_{\operatorname{Re} z \leq 0} |R(z)|. \quad (11.8)$$

Remark. This is a finite-dimensional version of a result of J. von Neumann (1951). A short proof is given in Hairer, Bader & Lubich (1982). The idea of the following proof is due to M. Crouzeix (unpublished).

Proof. a) Normal matrices can be transformed to diagonal form by a unitary matrix Q (see Exercise 3 of Section I.12). Hence, $A = QDQ^*$, where $D = \operatorname{diag}\{\lambda_1, \dots, \lambda_n\}$. In this case we have

$$\|R(A)\| = \|QR(D)Q^*\| = \|R(D)\| = \max_{i=1, \dots, n} |R(\lambda_i)|,$$

and (11.8) follows from (11.6), because the eigenvalues of A satisfy $\operatorname{Re} \lambda_i \leq 0$.

b) For a general A we consider the matrix function

$$A(\omega) = \frac{\omega}{2}(A + A^*) + \frac{1}{2}(A - A^*).$$

We see from the identity

$$\langle v, A(\omega)v \rangle = \omega \operatorname{Re} \langle v, Av \rangle + i \operatorname{Im} \langle v, Av \rangle$$

that $A(\omega)$ satisfies (11.6) for all ω with $\operatorname{Re} \omega \geq 0$, so that also the eigenvalues of $A(\omega)$ satisfy $\operatorname{Re} \lambda(\omega) \leq 0$ for $\operatorname{Re} \omega \geq 0$. Therefore, the rational function

$$\varphi(\omega) = \langle u, R(A(\omega))v \rangle$$

$(u, v$ fixed) has no poles in $\operatorname{Re} \omega \geq 0$. Using $A(1) = A$ we obtain from the maximum principle that

$$\begin{aligned} \langle u, R(A)v \rangle &= \varphi(1) \leq \sup_{y \in \mathbb{R}} \varphi(iy) \leq \sup_{y \in \mathbb{R}} \|R(A(iy))\| \|u\| \|v\| \\ &\leq \sup_{\operatorname{Re} z \leq 0} |R(z)| \|u\| \|v\|. \end{aligned} \quad (11.9)$$

The last inequality of (11.9) follows from part (a), because $A(iy)$ is a normal matrix (i.e., $A(iy)A(iy)^* = A(iy)^*A(iy)$). Formula (11.8) is now an immediate consequence of (11.9) and of the fact that $\|C\| = \sup_{\|u\| \leq 1, \|v\| \leq 1} \langle u, Cv \rangle$. \square

Corollary 11.3. *If the rational function $R(z)$ is A -stable, then the numerical solution $y_{n+1} = R(hA)y_n$ is contractive in the Euclidean norm (i.e., $\|y_{n+1}\| \leq \|y_n\|$), whenever (11.6) is satisfied.*

Proof. A -stability implies that $\max_{\operatorname{Re} z \leq 0} |R(z)| \leq 1$. \square

Corollary 11.4. *If a matrix A satisfies $\operatorname{Re} \langle v, Av \rangle \leq \nu \|v\|^2$ for all $v \in \mathbb{C}^n$, then*

$$\|R(A)\| \leq \sup_{\operatorname{Re} z \leq \nu} |R(z)|. \quad (11.10)$$

Proof. Apply Theorem 11.2 to $\tilde{R}(z) = R(z + \nu)$ and $\tilde{A} = A - \nu I$. \square

Error Growth Function for Linear Problems

Guided by the above estimate, we define

$$\varphi_R(x) := \sup_{\operatorname{Re} z \leq x} |R(z)|. \quad (11.11)$$

This function is called *error growth function* (for linear problems). It is continuous and monotonically increasing. If $R(z)$ is analytic in the half-plane $\operatorname{Re} z < x$, the maximum principle implies that

$$\varphi_R(x) = \sup_{y \in \mathbb{R}} |R(x + iy)|.$$

Examples.

1. Implicit Euler method:

$$R(z) = \frac{1}{1-z} \quad \varphi_R(x) = \begin{cases} R(x) & \text{if } -\infty < x < 1 \\ \infty & \text{if } 1 \leq x. \end{cases} \quad (11.12)$$

2. The stability function of the θ -method (or of a one-stage Rosenbrock method):

$$R(z) = \frac{1 + (1 - \theta)z}{1 - \theta z} \quad \varphi_R(x) = \begin{cases} |R(\infty)| & \text{if } x \leq \xi_0 \\ R(x) & \text{if } \xi_0 \leq x < 1/\theta \\ \infty & \text{if } 1/\theta \leq x, \end{cases} \quad (11.13)$$

where $\xi_0 = (1 - 2\theta)/(2\theta(1 - \theta))$ for $0 < \theta < 1$ and $\xi_0 = -\infty$ for $\theta \geq 1$.

3. The (0,2)-Padé approximation:

$$R(z) = \frac{1}{1 - z + z^2/2} \quad \varphi_R(x) = \begin{cases} R(x) & \text{if } -\infty < x \leq 0 \\ \frac{1}{1 - x} & \text{if } 0 \leq x < 1 \\ \infty & \text{if } 1 \leq x. \end{cases} \quad (11.14)$$

4. The (1,2)-Padé approximation $R(z) = \frac{1 + z/3}{1 - 2z/3 + z^2/6}$:

$$\varphi_R(x) = \begin{cases} |R(x)| & \text{if } -\infty < x \leq \xi_0 \\ \frac{\sqrt{3\sqrt{12x^2 + 12x + 9} + 10x + 7}}{2(2 - x)} & \text{if } \xi_0 \leq x < 2 \\ \infty & \text{if } 2 \leq x, \end{cases} \quad (11.15)$$

where $\xi_0 = -6 - 3\sqrt{10}$.

5. The (2,2)-Padé approximation $R(z) = \frac{1 + z/2 + z^2/12}{1 - z/2 + z^2/12}$:

$$\varphi_R(x) = \begin{cases} 1 & \text{if } -\infty < x \leq 0 \\ \frac{2x + \sqrt{9 + 3x^2}}{3 - x} & \text{if } 0 \leq x < 3 \\ \infty & \text{if } 3 \leq x. \end{cases} \quad (11.16)$$

The next two theorems give some general results on the shape of $\varphi_R(x)$.

Theorem 11.5. *Let $R(z)$ be an A -stable approximation to e^z of exact order p , i.e., $R(z) = e^z - Cz^{p+1} + \mathcal{O}(z^{p+2})$ with $C \neq 0$. If additionally $|R(iy)| < 1$ for $y \neq 0$ and $|R(\infty)| < 1$, then we have*

a) if p is odd

$$\varphi_R(x) = e^x + \mathcal{O}(x^{p+1}) \quad \text{for } x \rightarrow 0. \quad (11.17)$$

b) if p is even we have (11.17) only for $(-1)^{p/2}Cx > 0$, otherwise

$$\varphi_R(x) = e^x + \mathcal{O}(x^{r+1}) \quad \text{for } x \rightarrow 0 \quad (11.18)$$

for some positive rational number $r \leq p/2$.

Proof. The assumptions imply that for $x \rightarrow 0$ the maximum of $\{|R(x + iy)|; y \in \mathbb{R}\}$ must be located near the origin. We further observe that it must lie within the order star $A = \{z \in \mathbb{C}; |R(z)| > |e^z|\}$. If p is odd, the order star consists of $p + 1$ sectors near the origin (Lemma 4.3) and, asymptotically for $z \rightarrow \infty$, all elements of A satisfy $|z| \leq D|x|$, $D < \infty$. Therefore

$$|R(z)| = e^x + \mathcal{O}(|z|^{p+1}) = e^x + \mathcal{O}(x^{p+1}) \quad \text{for } x \rightarrow 0.$$

The same argument applies if p is even and $(-1)^{p/2}Cx > 0$. In the remaining case (p even and $(-1)^{p/2}Cx < 0$) the maximum of $\{|R(x + iy)|; y \in \mathbb{R}\}$ is attained near the imaginary axis and a more detailed analysis is necessary (Hairer, Bader & Lubich 1982). \square

Theorem 11.6 (Hairer & Zennaro 1996). *For an A -stable approximation to e^z the function $\varphi_R(x)$ is superexponential, i.e., it satisfies $\varphi_R(0) = 1$ and*

$$\varphi_R(x_1) \varphi_R(x_2) \leq \varphi_R(x_1 + x_2) \quad (11.19)$$

for all x_1, x_2 having the same sign.

Proof. A -stability is equivalent to $\varphi_R(0) = 1$. It therefore remains to verify (11.19). Let x_1 and x_2 be fixed (both ≤ 0 or both ≥ 0) and assume $\varphi_R(x_1 + x_2) < \infty$. The idea is to consider the rational function

$$S(z) = R(a - z)R(z)$$

where $a \in \mathbb{C}$ is a parameter satisfying $\operatorname{Re} a \leq x_1 + x_2$. Due to A -stability and $\varphi_R(x_1 + x_2) < \infty$, $S(z)$ is analytic on the stripe $0 \leq \operatorname{Re} z \leq x_1 + x_2$ (or $x_1 + x_2 \leq \operatorname{Re} z \leq 0$), and its modulus is bounded by $\varphi_R(x_1 + x_2)$ on the border. By the maximum principle we therefore have for all z in the considered stripe

$$|R(a - z)R(z)| \leq \varphi_R(x_1 + x_2).$$

We now choose z on the line $\operatorname{Re} z = x_2$ in such a way that $|R(z)|$ becomes maximal; then, we choose a on the line $\operatorname{Re} a = x_1 + x_2$ (i.e., $\operatorname{Re}(a - z) = x_1$) such that $|R(a - z)|$ becomes maximal (eventually one has to consider limits). This proves (11.19). \square

Property (11.19) has an interesting practical interpretation. Consider a numerical solution y_n obtained with variable step sizes. Repeated application of (11.4) and Corollary 11.4 implies

$$\|y_m\| \leq \left(\prod_{k=0}^{m-1} \varphi_R(h_k \mu) \right) \cdot \|y_0\|, \quad (11.20)$$

if the problem $y' = Ay$ satisfies (11.7) with $\mu = \mu_2(A)$. For $\mu < 0$ and for an A -stable method all factors $\varphi_R(h_k \mu)$ are smaller than one. If in addition $|R(\infty)| < 1$,

these factors are close to one only for $h_k \rightarrow 0$. The inequality (11.19), written as

$$\varphi_R(h_k \mu) \varphi_R(h_{k+1} \mu) \leq \varphi((h_k + h_{k+1}) \mu),$$

means that replacing two consecutive steps by one large step of size $h_k + h_{k+1}$ increases the upper bound (11.20). Therefore, after combining several consecutive steps (if necessary), we may assume $h_k \geq h > 0$ for all k . This implies that $\|y_m\| \leq \varrho^m \|y_0\|$ with $\varrho = \varphi_R(h\mu) < 1$. Hence, for any mesh x_0, x_1, \dots with $x_m \rightarrow \infty$, we have asymptotic stability, i.e., $\|y_m\| \rightarrow 0$ for $m \rightarrow \infty$. Under additional restrictions on the step size, sharper bounds on $\|y_m\|$ can be obtained (Exercise 3).

Small Nonlinear Perturbations

The above estimates, valid only for linear autonomous equations $y' = Jy$, can be extended to problems with small nonlinear perturbations, so-called *semi-linear* problems

$$y' = Jy + g(x, y) \quad (11.21)$$

where

$$\langle y, Jy \rangle \leq \mu \|y\|^2 \quad (11.22a)$$

$$\|g(x, y) - g(x, z)\| \leq L \|y - z\| \quad (11.22b)$$

with L assumed to be small.

Here, in the presence of nonlinearities, stability properties are obtained by estimating the *distance* of two neighbouring solutions $y(x)$ and $\hat{y}(x)$. Instead of (11.5) we therefore have

$$\frac{d}{dx} \|y(x) - \hat{y}(x)\|^2 = 2 \langle y' - \hat{y}', y - \hat{y} \rangle$$

which gives, after inserting (11.21) for y' and \hat{y}' , using the Cauchy-Schwarz inequality and the estimates (11.22)

$$\frac{d}{dx} \|y(x) - \hat{y}(x)\|^2 \leq 2(\mu + L) \|y(x) - \hat{y}(x)\|^2. \quad (11.23)$$

We thus have contractivity whenever $\mu + L \leq 0$.

We now want to establish the same property for the *numerical* solutions. In principle, these estimates can be carried out for all methods of this chapter; however, since the subsequent sections will deal with so many nice properties of implicit Runge-Kutta methods, we shall concentrate here on Rosenbrock methods.

Example 11.7. Consider the 1-stage Rosenbrock method

$$\begin{aligned} (I - \gamma h J) k_1 &= h f(x_0, y_0) \\ y_1 &= y_0 + k_1 \end{aligned} \quad (11.24)$$

with $\gamma > 0$ as a free parameter. Its stability function is

$$R(z) = \frac{1 + (1 - \gamma)z}{1 - \gamma z}$$

and we have A -stability for $\gamma \geq 1/2$. Application of (11.24) to (11.21) yields

$$y_1 = R(hJ)y_0 + (I - \gamma hJ)^{-1}hg(x_0, y_0). \quad (11.25)$$

From von Neumann's theorem (Corollary 11.4) we obtain $\|(I - \gamma hJ)^{-1}\| \leq (1 - \gamma h\mu)^{-1}$ and $\|R(hJ)\| \leq \varphi_R(h\mu)$ with φ_R given in (11.13). If we take a second numerical solution \hat{y}_1 , also defined by (11.25), its difference to y_1 can be estimated by

$$\|y_1 - \hat{y}_1\| \leq \left(R(h\mu) + \frac{hL}{1 - \gamma h\mu} \right) \|y_0 - \hat{y}_0\| = \left(1 + \frac{h(\mu + L)}{1 - \gamma h\mu} \right) \|y_0 - \hat{y}_0\|$$

whenever $\xi_0 < h\mu < 1/\gamma$ with ξ_0 given in (11.13). Therefore contractivity occurs for $\mu + L \leq 0$, as desired.

For the general Rosenbrock method (7.4) applied to problem (11.21)

$$\begin{aligned} k_i &= hg(x_0 + c_i h, u_i) + hJy_0 + hJ \sum_{j=1}^i (a_{ij} + \gamma_{ij}) k_j \\ u_i &= y_0 + \sum_{j=1}^{i-1} a_{ij} k_j, \quad y_1 = y_0 + \sum_{i=1}^s b_i k_i \end{aligned}$$

we easily find the following analogue of the variation of constants formula.

Theorem 11.8. *The numerical solution of a Rosenbrock method applied to (11.21) can be written as*

$$\begin{aligned} y_1 &= R(hJ)y_0 + h \sum_{i=1}^s b_i(hJ)g(x_0 + c_i h, u_i) \\ u_i &= R_i(hJ)y_0 + h \sum_{j=1}^{i-1} a_{ij}(hJ)g(x_0 + c_j h, u_j), \quad i = 1, \dots, s. \end{aligned} \quad (11.26)$$

Here $R(z)$ is the stability function, $R_i(z)$ are the so-called internal stability functions and $b_i(z)$, $a_{ij}(z)$ are rational functions whose only pole is $1/\gamma$ and which satisfy $b_i(\infty) = 0$, $a_{ij}(\infty) = 0$. \square

Remark. For many classes of linearly implicit methods (e.g., the methods of van der Houwen (1977), Friedli (1978), Strehmel & Weiner (1982), etc.), the numerical solution can be expressed by (11.26) with certain rational functions. Thus the following analysis can be applied to these methods as well.

We now take a second numerical solution $\widehat{y}_0, \widehat{u}_i, \widehat{y}_1$ (again defined by (11.26)), take the difference to y_1 and apply the triangle inequality. Using von Neumann's theorem (Corollary 11.4) the assumptions (11.22) then imply

$$\begin{aligned} \|\widehat{y}_1 - y_1\| &\leq \varphi_R(h\mu) \|\widehat{y}_0 - y_0\| + hL \sum_{i=1}^s \varphi_{b_i}(h\mu) \|\widehat{u}_i - u_i\| \\ \|\widehat{u}_i - u_i\| &\leq \varphi_{R_i}(h\mu) \|\widehat{y}_0 - y_0\| + hL \sum_{j=1}^{i-1} \varphi_{a_{ij}}(h\mu) \|e\widehat{u}_j - u_j\|. \end{aligned} \quad (11.27)$$

Inserting the second inequality of (11.27) repeatedly into the first one yields

Theorem 11.9. *Under the assumption (11.22) the difference of two numerical solutions of (7.4) can be estimated by*

$$\|\widehat{y}_1 - y_1\| \leq (\varphi_R(h\mu) + chL) \|\widehat{y}_0 - y_0\| \quad (11.28)$$

where $\varphi_R(x)$ is given by (11.11) ($R(z)$ is the stability function of (7.4)) and c is a constant depending smoothly on hL and $h\mu$ but not on $\|J\|$ (which represents the stiffness of the problem). \square

This estimate shows numerical contractivity whenever $\varphi_R(h\mu) + hL^* \leq 0$. In Theorem 11.5 we have shown under certain assumptions that $\varphi_R(x) = 1 + x + o(x)$, so contractivity holds essentially for $\mu + L^* \leq 0$. In any case we have that A -stability implies

$$\|\widehat{y}_1 - y_1\| \leq (1 + hC^*) \|\widehat{y}_0 - y_0\|$$

for $h\mu \leq \text{Const}$. Here, C^* is a constant independent of the stiffness of (11.21).

Remark. Since the rational functions b_i and a_{ij} in (11.26) vanish at infinity, also $(1 - \gamma hJ)b_i(hJ)$ and $(1 - \gamma hJ)a_{ij}(hJ)$ are uniformly bounded for J satisfying (11.22) and for $h\mu \leq C < \gamma^{-1}$. Instead of the second condition of (11.22) we may therefore require that

$$\|(I - \gamma hJ)^{-1} h(g(x, y) - g(x, z))\| \leq \ell \|y - z\|, \quad (11.29)$$

and the statement of Theorem 11.9 holds with hL replaced by ℓ . Observe that the assumption (11.22) implies (11.29) with $\ell = hL/(1 - \gamma h\mu)$. However, in some special situations the number ℓ may be significantly smaller than hL . Related techniques are used by Hundsdorfer (1985) and Strehmel & Weiner (1987) to prove contractivity and convergence for linearly implicit methods. Ostermann (1988) applies these ideas to nonlinear singular perturbation problems, where $hL = \mathcal{O}(h\varepsilon^{-1})$ with some very small ε ($\varepsilon \ll h$), but ℓ can be bounded independently of ε^{-1} .

Contractivity in $\|\cdot\|_\infty$ and $\|\cdot\|_1$

The study of contractivity in general norms has been carried out mainly by Spijker (1983, 1985) and his collaborators. Similar techniques of proof can be found in Bolley & Crouzeix (1978), where a related problem (monotonicity) is treated.

The following theorem gives a condition which is *necessary* for contractivity just for the special equation (11.2) and for one of the two norms $\|\cdot\|_\infty$ or $\|\cdot\|_1$. Later, the same condition will also turn out to be *sufficient* for general problems and *all* norms.

Theorem 11.10. *Let A be the n -dimensional matrix of (11.2) with fixed $\lambda \geq 0$. For a rational function $R(z)$ satisfying $R(0) = 1$ we have*

$$\|R(hA)\|_\infty \leq 1 \quad \text{in all dimensions } n = 1, 2, \dots \quad (11.30)$$

only if

$$R^{(j)}(x) \geq 0 \quad \text{for } x \in [-\lambda h, 0] \quad \text{and } j = 0, 1, 2, \dots \quad (11.31)$$

(The same statement is true, if $\|\cdot\|_\infty$ in (11.30) is replaced by $\|\cdot\|_1$).

Proof. We put $h = 1$ and write $A = -\lambda I + \lambda N$, where N is a nilpotent matrix. In a suitable norm, $\|N\|$ is arbitrarily small and therefore we have by Taylor expansion and $N^n = 0$

$$R(A) = \sum_{j=0}^{n-1} R^{(j)}(-\lambda) \frac{(\lambda N)^j}{j!}.$$

This means (e.g. for $n = 4$)

$$R(A) = \begin{pmatrix} R(-\lambda) & \lambda R'(-\lambda) & \frac{\lambda^2}{2!} R''(-\lambda) & \frac{\lambda^3}{3!} R'''(-\lambda) \\ & R(-\lambda) & \lambda R'(-\lambda) & \frac{\lambda^2}{2!} R''(-\lambda) \\ & & R(-\lambda) & \lambda R'(-\lambda) \\ & & & R(-\lambda) \end{pmatrix}.$$

Application of Formula (I.9.11') shows that $\|R(A)\|_\infty \leq 1$ (or $\|R(A)\|_1 \leq 1$) is equivalent to

$$\sum_{j=0}^{n-1} |R^{(j)}(-\lambda)| \frac{\lambda^j}{j!} \leq 1. \quad (11.32)$$

If (11.32) is valid for all $n \geq 1$, the series

$$\sum_{j \geq 0} R^{(j)}(-\lambda) \frac{\lambda^j}{j!} \quad (11.33)$$

is absolutely convergent, and therefore we have

$$1 = R(0) = \sum_{j \geq 0} R^{(j)}(-\lambda) \frac{\lambda^j}{j!} \leq \sum_{j \geq 0} |R^{(j)}(-\lambda)| \frac{\lambda^j}{j!} \leq 1$$

implying $R^{(j)}(-\lambda) \geq 0$ for all $j \geq 0$. Since the Taylor expansion

$$R^{(j)}(x) = \sum_{k \geq j} R^{(k)}(-\lambda) \frac{(x + \lambda)^{k-j}}{(k-j)!}$$

consists for $x \geq -\lambda$ only of non-negative terms, we have (11.31). \square

The next theorem shows that condition (11.31) is sufficient for contractivity in arbitrary norms. It can readily be applied to the system (11.2), since its matrix satisfies $\|A + \lambda I\|_\infty = \lambda$.

Theorem 11.11. *Consider an arbitrary norm and let A be such that for some $\lambda \geq 0$,*

$$\|A + \lambda I\| \leq \lambda. \quad (11.34)$$

If the stability function of a method satisfies $R(0) = 1$ and

$$R^{(j)}(x) \geq 0 \quad \text{for } x \in [-\varrho, 0] \quad \text{and } j = 0, 1, 2, \dots \quad (11.35)$$

then we have numerical contractivity $\|R(hA)\| \leq 1$, whenever $h\lambda \leq \varrho$.

Proof. We again put $h = 1$. Since for $0 \leq \lambda \leq \varrho$ we have $R^{(j)}(-\lambda) \geq 0$ for all j , the function

$$R(z) = \sum_{j \geq 0} R^{(j)}(-\lambda) \frac{(z + \lambda)^j}{j!} \quad (11.36)$$

satisfies $|R(z)| \leq R(-\lambda + r)$ for all complex z in the disk $|z + \lambda| \leq r$. This property and (11.35) imply that no pole of $R(z)$ can lie in $|z + \lambda| \leq \lambda$, so that the radius of convergence of (11.36) is strictly larger than λ . Consequently we have from (11.34)

$$R(A) = \sum_{j \geq 0} R^{(j)}(-\lambda) \frac{(A + \lambda I)^j}{j!}. \quad (11.37)$$

The triangle inequality applied to (11.37) yields the conclusion. \square

Study of the Threshold Factor

Definition 11.12. The largest ϱ satisfying (11.35) is called the *threshold-factor* of $R(z)$.

Example 11.13. The implicit Euler method, for which

$$R^{(j)}(x) = \frac{j!}{(1-x)^{j+1}}, \quad j = 0, 1, 2, \dots,$$

satisfies (11.35) for all $\varrho > 0$. It possesses a threshold-factor $\varrho = \infty$.

Example 11.14 (Threshold-factor for Padé-approximations). The derivatives of the polynomials

$$R_{k0}(z) = 1 + z + \frac{z^2}{2!} + \dots + \frac{z^k}{k!}$$

are easily calculated; the most dangerous one is $1 + z$, therefore $\varrho = 1$ for all k .

The Padé approximations $R_{k1}(z)$ possess one simple pole only, so they can be written in the form

$$R_{k1}(z) = \frac{a}{1 - bz} + \text{polynomial in } z,$$

which has only a finite number of derivatives which can change sign (see Example 11.13). The numerical values obtained are shown in Table 11.1.

The functions $R_{k2}(z)$ possess no real pole (see Sect. IV.4). But the property $|R(z)| \leq R(-\varrho + r)$ for $|z + \varrho| \leq r$ (see proof of Theorem 11.10) means that the maximum of $|R(z)|$ on the circle with center $-\varrho$ and radius r is assumed to the right on the real axis. For increasing r , the first pole met by this circle must therefore be real and to the right of $-\varrho$. This is not possible here and therefore the approximations $R_{k2}(z)$ *never* satisfy property (11.35). This is indicated by an asterisk (*) in Table 11.1.

All further values of Table 11.1 were computed using the decomposition of $R(z)$ into partial fractions and are cited from Kraaijevanger (1986) and van de Griend & Kraaijevanger (1986).

Table 11.1. Threshold-factors of Padé approximations

k	0	1	2	3	4	5	6
$j = 0$	—	1	1	1	1	1	1
$j = 1$	∞	2	2.196	2.350	2.477	2.586	2.682
$j = 2$	*	*	*	*	*	*	*
$j = 3$	0.584	1.195	1.703	2.208	2.710	3.212	3.713
$j = 4$	*	*	*	*	*	*	*
$j = 5$	0.353	0.770	1.081	1.424	1.794	2.185	2.590

It is curious to observe that in this table the methods with the largest threshold-factors are precisely those which are not A -stable. An exception is the implicit Euler method ($k = 0, j = 1$) for which $\varrho = \infty$.

Absolutely Monotonic Functions

... on peut définir la fonction e^x comme la seule fonction absolument monotone sur tout le demi-axe négatif qui prend à l'origine, ainsi que sa dérivée première [*sic*] la valeur un.

(S. Bernstein 1928)

A thorough study of real functions satisfying (11.35) was begun by S. Bernstein (1914) and continued by F. Hausdorff (1921). Such functions are called *absolutely monotonic* in $[-\varrho, 0]$. Later, S. Bernstein (1928) gave the following characterization of functions which are absolutely monotonic in $(-\infty, 0]$ (see also D.V. Widder 1946).

Theorem 11.15 (Bernstein 1928). *A necessary and sufficient condition that $R(x)$ be absolutely monotonic in $(-\infty, 0]$ is that*

$$R(x) = \int_0^\infty e^{xt} d\alpha(t), \quad (11.38)$$

where $\alpha(t)$ is bounded and non-decreasing and the integral converges for $-\infty < x \leq 0$.

This is a hard result and the main key for the next two theorems. It does not seem to permit an elementary and easy proof. We therefore refer to the original literature, S. Bernstein (1928). For a more recent description see e.g. Widder (1946), p. 160. From this result we immediately get the “limit case $\lambda \rightarrow \infty$ ” of Theorem 11.11, which also holds for an arbitrary norm.

Theorem 11.16. *Let $R(x)$ be absolutely monotonic in $(-\infty, 0]$, $R(0) = 1$ and A a matrix with non-positive logarithmic norm $\mu(A) \leq 0$, then*

$$\|R(A)\| \leq 1.$$

Proof. By Theorem I.10.6 we have for the solution $y(x) = e^{Ax}y_0$ of $y' = Ay$ that $\|y(x)\| \leq \|y_0\|$, hence also $\|e^{Ax}\| \leq 1$ for $x \geq 0$. The statement now follows from

$$\|R(A)\| = \left\| \int_0^\infty e^{At} d\alpha(t) \right\| \leq \int_0^\infty \|e^{At}\| d\alpha(t) \leq \int_0^\infty d\alpha(t) = R(0) = 1$$

since $\alpha(t)$ is non-decreasing. □

The following result proves that no Runge-Kutta method of order $p > 1$ can have a stability function which is absolutely monotonic in $(-\infty, 0]$.

Theorem 11.17. *If $R(x)$ is absolutely monotonic in $(-\infty, 0]$ and*

$$R(x) = 1 + x + x^2/2 + \mathcal{O}(x^3) \quad \text{for } x \rightarrow 0,$$

then $R(x) = e^x$.

Proof (Bolley & Crouzeix 1978). It follows from (11.38) that

$$R^{(j)}(0) = \int_0^\infty t^j d\alpha(t).$$

Since $R(0) = R'(0) = R''(0) = 1$, this yields

$$\int_0^\infty (1-t)^2 d\alpha(t) = 0.$$

Consequently, $\alpha(t)$ must be the Heaviside function ($\alpha(t) = 0$ for $t \leq 1$ and $\alpha(t) = 1$ for $t > 1$). Inserted into (11.38) this gives $R(x) = e^x$. \square

Exercises

1. Prove Formula (11.14). For given x , study the set of y -values for which $|R(x + iy)|$ attains its maximum.
2. Show that the error growth function (11.11) for an A -stable $R(z)$ of order $p \geq 1$ satisfies

$$\varphi_R(x) > e^x \quad \text{for all } x \neq 0.$$

Hint. You can study the order star on parallel lines $\{x + iy, y \in \mathbb{R}\}$ (Hairer, Bader & Lubich 1982), or you can use the fact that $\varphi_R(x)$ is superexponential.

3. (Hairer & Zennaro 1996). Let $|R(\infty)| < 1$ and consider a mesh x_0, x_1, \dots with step sizes $h_k = x_{k+1} - x_k$ satisfying $h_{k+1} \leq ch_k$ ($c > 1$). Prove the existence of constants $C > 0$ and $\alpha > 0$ such that

$$\|y_m\| \leq C(x_m - x_0)^{-\alpha} \|y_0\| \quad \text{for } m = 1, 2, \dots$$

4. (Kraaijevanger 1986). Let $R(z)$ be a polynomial of degree s satisfying $R(z) = e^z + \mathcal{O}(z^{p+1})$. Then the threshold factor ϱ (Definition 11.11) is restricted by

$$\varrho \leq s - p + 1.$$

Hint. Justify the formula

$$R^{(p-1)}(z) = \sum_{j=0}^{s-p+1} \alpha_j \left(1 + \frac{z}{\varrho}\right)^j, \quad \alpha_j \geq 0$$

and deduce the result from $R^{(p-1)}(0) = R^{(p)}(0) = 1$.

5. Let ϱ be the threshold factor of the rational function $R(z)$. Show that its stability domain contains the disc $|z + \varrho| \leq \varrho$.

IV.12 B-Stability and Contractivity

Next we need a generalization of the notion of A -stability. The most natural generalization would be to consider the case that $x(t)$ is a uniform-asymptotically stable solution ... in the sense of Liapunov theory ... but this case seems to be a little too wide.

(G. Dahlquist 1963)

The theoretical analysis of the application of numerical methods on stiff nonlinear problems is still fairly incomplete.

(G. Dahlquist 1975)

Here we enter a new era, the study of stability and convergence for general *non-linear* systems. All the “crimes” and diverse omissions of which we have been guilty in earlier sections, especially in Sect. IV.2, shall now be repaired.

Large parts of Dahlquist’s (1963) paper deal with a generalization of A -stability to nonlinear problems. His search for a sufficiently general class of nonlinear systems was finally successful 12 years later. In his talk at the Dundee conference of July 1975 he proposed to consider differential equations satisfying a one-sided Lipschitz condition, and he presented some first results for multistep methods. J.C. Butcher (1975) then extended (on the flight back from the conference) the ideas to implicit Runge-Kutta methods and the concept of B -stability was born.

One-Sided Lipschitz Condition

We consider the nonlinear differential equation

$$y' = f(x, y) \quad (12.1)$$

such that for the Euclidean norm the *one-sided Lipschitz condition*

$$\langle f(x, y) - f(x, z), y - z \rangle \leq \nu \|y - z\|^2 \quad (12.2)$$

holds. The number ν is the *one-sided Lipschitz constant* of f . This definition is motivated by the following result.

Lemma 12.1. *Let $f(x, y)$ be continuous and satisfy (12.2). Then, for any two solutions $y(x)$ and $z(x)$ of (12.1) we have*

$$\|y(x) - z(x)\| \leq \|y(x_0) - z(x_0)\| \cdot e^{\nu(x-x_0)} \quad \text{for } x \geq x_0.$$

Proof. Differentiation of $m(x) = \|y(x) - z(x)\|^2$ yields

$$m'(x) = 2\langle f(x, y(x)) - f(x, z(x)), y(x) - z(x) \rangle \leq 2\nu m(x).$$

This differential inequality can be solved to give (see Theorem I.10.3)

$$m(x) \leq m(x_0)e^{2\nu(x-x_0)} \quad \text{for } x \geq x_0,$$

which is equivalent to the statement. □

Remarks. a) In an open convex set, condition (12.2) is equivalent to $\mu(\frac{\partial f}{\partial y}) \leq \nu$ (see Sect. I.10, Exercise 6), if f is continuously differentiable. Lemma 12.1 then becomes a special case of Theorem I.10.6.

b) For complex-valued y and f condition (12.2) has to be replaced by

$$\operatorname{Re} \langle f(x, y) - f(x, z), y - z \rangle \leq \nu \|y - z\|^2, \quad y, z \in \mathbb{C}^n, \quad (12.2')$$

and Lemma 12.1 remains valid.

B -Stability and Algebraic Stability

Whenever $\nu \leq 0$ in (12.2) the distance between any two solutions of (12.1) is a non-increasing function of x . The same property is then also desirable for the numerical solutions. We consider here implicit Runge-Kutta methods

$$y_1 = y_0 + h \sum_{i=1}^s b_i f(x_0 + c_i h, g_i), \quad (12.3a)$$

$$g_i = y_0 + h \sum_{j=1}^s a_{ij} f(x_0 + c_j h, g_j), \quad i = 1, \dots, s. \quad (12.3b)$$

Definition 12.2 (Butcher 1975). A Runge-Kutta method is called B -stable, if the contractivity condition

$$\langle f(x, y) - f(x, z), y - z \rangle \leq 0 \quad (12.2'')$$

implies for all $h \geq 0$

$$\|y_1 - \hat{y}_1\| \leq \|y_0 - \hat{y}_0\|.$$

Here, y_1 and \hat{y}_1 are the numerical approximations after one step starting with initial values y_0 and \hat{y}_0 , respectively.

Clearly, B -stability implies A -stability. This is seen by applying the above definition to $y' = \lambda y$, $\lambda \in \mathbb{C}$ or, more precisely, to

$$\begin{pmatrix} y'_1 \\ y'_2 \end{pmatrix} = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}. \quad (12.4)$$

Example 12.3. For the collocation methods based on Gaussian quadrature a simple proof of B -stability is possible (Wanner 1976). We denote by $u(x)$ and $\hat{u}(x)$ the collocation polynomials (see Definition II.7.6) for the initial values y_0 and \hat{y}_0 and differentiate the function $m(x) = \|u(x) - \hat{u}(x)\|^2$. At the collocation points $\xi_i = x_0 + c_i h$ we obtain

$$m'(\xi_i) = 2 \langle f(\xi_i, u(\xi_i)) - f(\xi_i, \hat{u}(\xi_i)), u(\xi_i) - \hat{u}(\xi_i) \rangle \leq 0.$$

The result then follows from the fact that Gaussian quadrature integrates the polynomial $m'(x)$ (which is of degree $2s-1$) exactly and that the weights b_i are positive:

$$\begin{aligned}\|y_1 - \hat{y}_1\|^2 &= m(x_0 + h) = m(x_0) + \int_{x_0}^{x_0+h} m'(x) dx \\ &= m(x_0) + h \sum_{i=1}^s b_i m'(x_0 + c_i h) \leq m(x_0) = \|y_0 - \hat{y}_0\|^2.\end{aligned}$$

An *algebraic criterion* for B -stability was found independently by Burrage & Butcher (1979) and Crouzeix (1979). The result is

Theorem 12.4. *If the coefficients of a Runge-Kutta method (12.3) satisfy*

- i) $b_i \geq 0$ for $i = 1, \dots, s$,
 - ii) $M = (m_{ij}) = (b_i a_{ij} + b_j a_{ji} - b_i b_j)_{i,j=1}^s$ *is non-negative definite,*
- then the method is B -stable.*

Definition 12.5. A Runge-Kutta method, satisfying (i) and (ii) of Theorem 12.4, is called *algebraically stable*.

Proof of Theorem 12.4. We introduce the differences

$$\begin{aligned}\Delta y_0 &= y_0 - \hat{y}_0, & \Delta y_1 &= y_1 - \hat{y}_1, & \Delta g_i &= g_i - \hat{g}_i, \\ \Delta f_i &= h(f(x_0 + c_i h, g_i) - f(x_0 + c_i h, \hat{g}_i)),\end{aligned}$$

and subtract the Runge-Kutta formulas (12.3) for y and \hat{y}

$$\Delta y_1 = \Delta y_0 + \sum_{i=1}^s b_i \Delta f_i, \quad (12.5a)$$

$$\Delta g_i = \Delta y_0 + \sum_{j=1}^s a_{ij} \Delta f_j. \quad (12.5b)$$

Next we take the square of Formula (12.5a)

$$\|\Delta y_1\|^2 = \|\Delta y_0\|^2 + 2 \sum_{i=1}^s b_i \langle \Delta f_i, \Delta y_0 \rangle + \sum_{i=1}^s \sum_{j=1}^s b_i b_j \langle \Delta f_i, \Delta f_j \rangle. \quad (12.6)$$

The main idea of the proof is now to compute Δy_0 from (12.5b) and insert this into (12.6). This gives

$$\|\Delta y_1\|^2 = \|\Delta y_0\|^2 + 2 \sum_{i=1}^s b_i \langle \Delta f_i, \Delta g_i \rangle - \sum_{i=1}^s \sum_{j=1}^s m_{ij} \langle \Delta f_i, \Delta f_j \rangle. \quad (12.7)$$

The statement now follows from the fact that $\langle \Delta f_i, \Delta g_i \rangle \leq 0$ by (12.2'') and that $\sum_{i,j=1}^s m_{ij} \langle \Delta f_i, \Delta f_j \rangle \geq 0$ (see Exercise 2). \square

Example 12.6. For the SDIRK method of Table 7.2 (Chapter II) the weights b_i are seen to be positive and the matrix M becomes

$$M = (\gamma - 1/4) \cdot \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

For $\gamma \geq 1/4$ this matrix is non-negative definite and we have B -stability. Exactly the same condition was obtained by studying its A -stability (c.f. (3.10)).

Some Algebraically Stable IRK Methods

La première de ces propriétés consiste en ce que tous les A_k sont positifs. (T.-J. Stieltjes 1884)

The general study of algebraic stability falls naturally into two steps: the positivity of the quadrature weights and the nonnegative-definiteness of the matrix M .

Theorem 12.7. Consider a quadrature formula $(c_i, b_i)_{i=1}^s$ of order p .

- a) If $p \geq 2s - 1$ then $b_i > 0$ for all i .
- b) If c_i are the zeros of (5.3) (Lobatto quadrature) then $b_i > 0$ for all i .

Proof (Stieltjes 1884). The first statement follows from the fact that for $p \geq 2s - 1$ polynomials of degree $2s - 2$ are integrated exactly, hence

$$b_i = \int_0^1 \prod_{j \neq i} \left(\frac{x - c_j}{c_i - c_j} \right)^2 dx > 0. \quad (12.8)$$

In the case of the Lobatto quadrature ($c_1 = 0$, $c_s = 1$ and $p = 2s - 2$) the factors for the indices $j = 1$ and $j = s$ are taken without squaring and the same argument applies. \square

In order to verify condition (ii) of Theorem 12.4 we find it convenient to use the W -transformation of Sect. IV.5 and to consider $W^T M W$ instead of M . In vector notation ($b = (b_1, \dots, b_s)^T$, $B = \text{diag}(b_1, \dots, b_s)$, $A = (a_{ij})$) we have

$$M = BA + A^T B - bb^T. \quad (12.9)$$

If we choose W according to Lemma 5.12, then $W^T B W = I$ and, since $W^T b = e_1 = (1, 0, \dots, 0)^T$, condition (ii) becomes equivalent to

$$W^T M W = X + X^T - e_1 e_1^T \quad \text{is non-negative definite} \quad (12.10)$$

where $X = W^{-1} A W = W^T B A W$ as in Theorem 5.11.

Theorem 12.8. Suppose that a Runge-Kutta method with distinct c_i and positive b_i satisfies the simplifying assumptions $B(2s - 2)$, $C(s - 1)$, $D(s - 1)$ (see beginning of Sect. IV.5). Then the method is algebraically stable if and only if $|R(\infty)| \leq 1$ (where $R(z)$ denotes the stability function).

Proof. Since the order of the quadrature formula is $p \geq 2s - 2$, the matrix W of Lemma 5.12 is

$$W = W_G D, \quad D = \text{diag}(1, \dots, 1, \alpha^{-1}), \quad (12.11)$$

where $W_G = (P_{j-1}(c_i))_{i,j=1}^s$ is as in (5.21), and $\alpha^2 = \sum_{i=1}^s b_i P_{s-1}^2(c_i) \neq 0$. Using the relation (observe that $W^T B W = I$)

$$X = W^{-1} A W = D^{-1} W_G^{-1} A W_G D = D W_G^T B A (W_G^T B)^{-1} D^{-1}$$

and applying Lemma 5.7 with $\eta = s - 1$ and Lemma 5.8 with $\xi = s - 1$, we obtain

$$X = \begin{pmatrix} 1/2 & -\xi_1 & & & \\ \xi_1 & 0 & & & \\ & & \ddots & & \\ & & & -\xi_{s-2} & \\ & & & \xi_{s-2} & 0 & -\alpha \xi_{s-1} \\ & & & & \alpha \xi_{s-1} & \beta \end{pmatrix}.$$

If this matrix is inserted into (12.10) then, marvellous surprise, everything cancels with the exception of β . Therefore, condition (ii) of Theorem 12.4 is equivalent to $\beta \geq 0$.

Using the representation (5.31) of the stability function we obtain by developing the determinants

$$|R(\infty)| = \left| \frac{\det(X - e_1 e_1^T)}{\det X} \right| = \left| \frac{\beta d_{s-1} - \alpha^2 \xi_{s-1}^2 d_{s-2}}{\beta d_{s-1} + \alpha^2 \xi_{s-1}^2 d_{s-2}} \right|, \quad (12.12)$$

where $d_k = k!/(2k)!$ is the determinant of the k -dimensional matrix X_G of (5.13). Since $\alpha^2 \xi_{s-1}^2 d_{s-2} > 0$, the expression (12.12) is bounded by 1 iff $\beta \geq 0$. This proves the statement. \square

Comparing these theorems with Table 5.13 yields

Theorem 12.9. *The methods Gauss, Radau IA, Radau IIA and Lobatto IIIC are algebraically stable and therefore also B-stable.* \square

AN-Stability

A-stability theory is based on the autonomous linear equation $y' = \lambda y$, whereas B-stability is based on general nonlinear systems $y' = f(x, y)$. The question arises whether there is a reasonable stability theory *between* these two extremes. A natural approach would be to study the scalar, linear, nonautonomous equation

$$y' = \lambda(x)y, \quad \text{Re } \lambda(x) \leq 0, \quad (12.13)$$

where $\lambda(x)$ is an arbitrarily varying complex-valued function (Burrage & Butcher 1979, Scherer 1979). The somewhat surprising result of this subsection will be that stability for (12.13) will, for most RK-methods, be equivalent to B-stability.

For the problem (12.13) the Runge-Kutta method (12.3) becomes (in vector notation $g = (g_1, \dots, g_s)^T$, $\mathbb{1} = (1, \dots, 1)^T$)

$$g = \mathbb{1}y_0 + AZg, \quad Z = \text{diag}(z_1, \dots, z_s), \quad z_j = h\lambda(x_0 + c_j h). \quad (12.14)$$

Computing g from (12.14) and inserting into (12.3a) gives

$$y_1 = K(Z)y_0, \quad K(Z) = 1 + b^T Z(I - AZ)^{-1} \mathbb{1}. \quad (12.15)$$

Definition 12.10. A Runge-Kutta method is called AN -stable, if

$$|K(Z)| \leq 1 \quad \begin{cases} \text{for all } Z = \text{diag}(z_1, \dots, z_s) \text{ satisfying } \text{Re } z_j \leq 0 \\ \text{and } z_j = z_k \text{ whenever } c_j = c_k \text{ } (j, k = 1, \dots, s). \end{cases}$$

Comparing (12.15) with (3.2) we find that

$$K(\text{diag}(z, z, \dots, z)) = R(z), \quad (12.16)$$

the usual stability function. Further, arguing as with (12.4), B -stability implies AN -stability. Therefore we have:

Theorem 12.11. *For Runge-Kutta methods it holds*

$$B\text{-stable} \Rightarrow AN\text{-stable} \Rightarrow A\text{-stable}. \quad \square$$

For the trapezoidal rule $y_1 = y_0 + \frac{h}{2}(f(x_0, y_0) + f(x_1, y_1))$ the function $K(Z)$ of (12.15) is given by

$$K(Z) = \frac{1 + z_1/2}{1 - z_2/2}. \quad (12.17)$$

Putting $z_2 = 0$ and $z_1 \rightarrow -\infty$ we see that this method is not AN -stable. More generally we have the following result.

Theorem 12.12 (Scherer 1979). *The Lobatto IIIA and Lobatto IIIB methods are not AN -stable and therefore not B -stable.*

Proof. As in Proposition 3.2 we find that

$$K(Z) = \frac{\det(I - (A - \mathbb{1}b^T)Z)}{\det(I - AZ)}. \quad (12.18)$$

By definition, the first row of A and the last row of $A - \mathbb{1}b^T$ vanish for the Lobatto IIIA methods (compare also the proof of Theorem 5.5). Therefore the denominator of $K(Z)$ does not depend on z_1 and the numerator not on z_s . If we put for example $z_2 = \dots = z_s = 0$, the function $K(Z)$ is unbounded for $z_1 \rightarrow -\infty$. This contradicts AN -stability.

For the Lobatto IIIB methods, one uses in a similar way that the last column of A and the first column of $A - \mathbb{1}b^T$ vanish. \square

The following result shows, as mentioned above, that AN -stability is much closer to B -stability than to A -stability.

Theorem 12.13 (Burrage & Butcher 1979). *Suppose that*

$$|K(Z)| \leq 1 \quad \left\{ \begin{array}{l} \text{for all } Z = \text{diag}(z_1, \dots, z_s) \text{ with } \text{Re } z_j \leq 0 \\ \text{and } |z_j| \leq \varepsilon \text{ for some } \varepsilon > 0, \end{array} \right. \quad (12.19)$$

then the method is algebraically stable (and hence also B -stable).

Proof. For $\Delta f_i := z_i \Delta g_i$ and $\Delta y_0 = 1$ the result of (12.5) is $\Delta y_1 = K(Z)$. Taking care of the fact that z_i need not be real, the computation of the proof of Theorem 12.4 shows that

$$|K(Z)|^2 - 1 = 2 \sum_{i=1}^s b_i \text{Re } z_i |g_i|^2 - \sum_{i,j=1}^s m_{ij} \bar{z}_i \bar{g}_i z_j g_j. \quad (12.20)$$

Here $g = (g_1, \dots, g_s)^T$ is a solution of (12.14) with $y_0 = 1$.

To prove that $b_i \geq 0$, choose $z_i = -\varepsilon < 0$ and $z_j = 0$ for $j \neq i$. Assumption (12.19) together with (12.20) implies

$$-2\varepsilon b_i |g_i|^2 - m_{ii} \varepsilon^2 |g_i|^2 \leq 0. \quad (12.21)$$

For sufficiently small ε , g_i is close to 1 and the second term in (12.21) is negligible for $b_i \neq 0$. Therefore, b_i must be non-negative.

To verify the second condition of algebraic stability we choose the purely imaginary numbers $z_j = i\varepsilon \xi_j$ ($\xi_j \in \mathbb{R}$). Since again $g_i = 1 + \mathcal{O}(\varepsilon)$ for $\varepsilon \rightarrow 0$, we have from (12.20) that

$$-\varepsilon^2 \sum_{i,j=1}^s m_{ij} \xi_i \xi_j + \mathcal{O}(\varepsilon^3) \leq 0.$$

Therefore, $M = (m_{ij})$ has to be non-negative definite. \square

Combining this result with those of Theorems 12.4 and 12.11 we obtain

Corollary 12.14. *For non-confluent Runge-Kutta methods (i.e., methods with all c_j distinct) the concepts of AN -stability, B -stability and algebraic stability are equivalent. \square*

An equivalence result (between B - and algebraic stability) for *confluent* Runge-Kutta methods is much more difficult to prove (see Theorem 12.18 below) and will be our next goal. To this end we first have to discuss *reducible* methods.

Reducible Runge-Kutta Methods

For an RK-method (12.3) it may happen that for all differential equations (12.1)

- i) some stages don't influence the numerical solution;
- ii) several g_i are identical.

In both situations the Runge-Kutta method can be simplified to an "equivalent" one with fewer stages.

For an illustration of situation (i) consider the method of Table 12.1. Its numerical solution is independent of g_2 and equivalent to the implicit Euler solution. For the method of Table 12.2 one easily verifies that $g_1 = g_2$, whenever the system (12.3b) possesses a unique solution. The method is thus equivalent to the implicit mid-point rule.

Table 12.1.

DJ -reducible method

1	1	0
1/2	1/4	1/4
	1	0

Table 12.2.

S -reducible method

1/2	1/2	0
1/2	1/4	1/4
	1/2	1/2

The situation (i) above can be made more precise as follows:

Definition 12.15 (Dahlquist & Jeltsch 1979). A Runge-Kutta method is called *DJ-reducible*, if for some non-empty index set $T \subset \{1, \dots, s\}$,

$$b_j = 0 \quad \text{for } j \in T \quad \text{and} \quad a_{ij} = 0 \quad \text{for } i \notin T, j \in T. \quad (12.22)$$

Otherwise it is called *DJ-irreducible*.

Condition (12.22) implies that the stages $j \in T$ don't influence the numerical solution. This is best seen by permuting the stages so that the elements of T are the last ones (Cooper 1985). Then the Runge-Kutta tableau becomes that of Table 12.3.

Table 12.3. DJ -reducibility

$$\begin{array}{c|cc} c_1 & A_{11} & 0 \\ c_2 & A_{21} & A_{22} \\ \hline & b_1^T & 0 \end{array} \quad \Rightarrow \quad \begin{array}{c|c} c_1 & A_{11} \\ \hline & b_1^T \end{array}$$

An interesting property of DJ -irreducible and algebraically stable Runge-Kutta methods was discovered by Dahlquist & Jeltsch (1979).

Theorem 12.16. A DJ -irreducible, algebraically stable Runge-Kutta method satisfies

$$b_i > 0 \quad \text{for } i = 1, \dots, s.$$

Proof. Suppose $b_j = 0$ for some index j . Then $m_{jj} = 0$ by definition of M . Since M is non-negative definite, all elements in the j th column of M must vanish (Exercise 11) so that $b_i a_{ij} = 0$ for all i . This implies (12.22) for the set $T = \{j | b_j = 0\}$, a contradiction to DJ -irreducibility. \square

An algebraic criterion for the situation (ii) was given for the first time (but incompletely) by Stetter (1973, p. 127) and finally by Hundsdorfer & Spijker (1981), see also Butcher (1987), p. 319, and Dekker & Verwer (1984), p. 108.

Definition 12.17. A Runge-Kutta method is S -reducible, if for some partition (S_1, \dots, S_r) of $\{1, \dots, s\}$ with $r < s$ we have for all l and m

$$\sum_{k \in S_m} a_{ik} = \sum_{k \in S_m} a_{jk} \quad \text{if } i, j \in S_l. \quad (12.23)$$

Otherwise it is called S -irreducible. Methods which are neither DJ -reducible nor S -reducible are called *irreducible*.

In order to understand condition (12.23) we assume that, after a certain permutation of the stages, $l \in S_l$ for $l = 1, \dots, r$. We then consider the r -stage method with coefficients

$$c_i^* = c_i, \quad a_{ij}^* = \sum_{k \in S_j} a_{ik}, \quad b_j^* = \sum_{k \in S_j} b_k. \quad (12.24)$$

Application of this new method to (12.1) yields $g_1^*, \dots, g_r^*, y_1^*$ and one easily verifies that g_i and y_1 defined by

$$g_i = g_l^* \quad \text{if } i \in S_l, \quad y_1 = y_1^*,$$

are a solution of the original method (12.3). For the method of Table 12.2 we have $S_1 = \{1, 2\}$. A further example of an S -reducible method is given in Table 12.4 of Sect. II.12 ($S_1 = \{1, 2, 3\}$ and $S_2 = \{4\}$).

The Equivalence Theorem for S -Irreducible Methods

Theorem 12.18 (Hundsdorfer & Spijker 1981). *For S -irreducible Runge-Kutta methods,*

$$B\text{-stable} \iff \text{algebraically stable}.$$

Proof. Because of Corollary 12.14, which covers nearly all cases of practical importance — and which was much easier to prove — this theorem seems to be of little practical interest. However, it is an interesting result which had been conjectured by many people for many years, so we reproduce its proof, which also includes the three Lemmas 12.19–12.21. The counter example of Exercise 6 below shows that S -irreducibility is a necessary hypothesis.

By Theorem 12.4 it is sufficient to prove that B -stability and S -irreducibility imply algebraic stability. For this we take s complex numbers z_1, \dots, z_s which satisfy $\operatorname{Re} z_j < 0$ and $|z_j| \leq \varepsilon$ for some sufficiently small $\varepsilon > 0$. We show that there exists a continuous function $f: \mathbb{C} \rightarrow \mathbb{C}$ satisfying

$$\operatorname{Re} \langle f(u) - f(v), u - v \rangle \leq 0 \quad \text{for all } u, v \in \mathbb{C}, \quad (12.25)$$

such that the Runge-Kutta solutions y_1, g_i and \hat{y}_1, \hat{g}_i corresponding to $y_0 = 0$, $\hat{y}_0 = 1$, $h = 1$ satisfy

$$f(\hat{g}_i) - f(g_i) = z_i(\hat{g}_i - g_i). \quad (12.26)$$

This yields $\hat{y}_1 - y_1 = K(Z)$ with $K(Z)$ given by (12.15). B -stability then implies $|K(Z)| \leq 1$. By continuity of $K(Z)$ near the origin we then have $|K(Z)| \leq 1$ for all z_j which satisfy $\operatorname{Re} z_j \leq 0$ and $|z_j| \leq \varepsilon$, so that Theorem 12.13 proves the statement.

Construction of the function f : we denote by Δg_i the solution of

$$\Delta g_i = 1 + \sum_{j=1}^s a_{ij} z_j \Delta g_j$$

(the solution exists uniquely if $|z_j| \leq \varepsilon$ and ε is sufficiently small). With ξ, η given by Lemma 12.19 (below) we define

$$\begin{aligned} g_i &= t\eta_i, & f(g_i) &= t\xi_i \\ \hat{g}_i &= g_i + \Delta g_i, & f(\hat{g}_i) &= f(g_i) + z_i \Delta g_i. \end{aligned} \quad (12.27)$$

This is well-defined for sufficiently large t (to be fixed later), because the η_i are distinct. Clearly, g_i and \hat{g}_i represent a Runge-Kutta solution for $y_0 = 0$ and $\hat{y}_0 = 1$, and (12.26) is satisfied by definition.

We next show that

$$\operatorname{Re} \langle f(u) - f(v), u - v \rangle < 0 \quad \text{if} \quad u \neq v \quad (12.28)$$

is satisfied for $u, v \in D = \{g_1, \dots, g_s, \hat{g}_1, \dots, \hat{g}_s\}$. This follows from the construction of ξ, η , if $u, v \in \{g_1, \dots, g_s\}$. If $u = g_i$ and $v = \hat{g}_i$ this is a consequence of (12.26). For the remaining case $u = \hat{g}_i, v \in D \setminus \{g_i, \hat{g}_i\}$ we have

$$\langle f(u) - f(v), u - v \rangle = t^2(\xi_i - \xi_j)(\eta_i - \eta_j) + \mathcal{O}(t) \quad \text{for } t \rightarrow \infty,$$

so that (12.28) is satisfied, if t is sufficiently large. Applying Lemma 12.20 below we find a continuous function $f: \mathbb{C} \rightarrow \mathbb{C}$ that extends (12.27) and satisfies (12.25). \square

To complete the above proof we still need the following three lemmas:

Lemma 12.19. *Let A be the coefficient matrix of an S -irreducible Runge-Kutta method. Then there exist vectors $\xi \in \mathbb{R}^s$ and $\eta = A\xi$ such that*

$$(\xi_i - \xi_j)(\eta_i - \eta_j) < 0 \quad \text{for } i \neq j. \quad (12.29)$$

Proof (see Butcher 1982). The first idea is to put

$$\xi = \mathbb{1} - \varepsilon A \mathbb{1} \quad \text{with} \quad \mathbb{1} = (1, 1, \dots, 1)^T,$$

so that η becomes

$$\eta = A\xi = A\mathbb{1} - \varepsilon A^2 \mathbb{1}.$$

If $c_i \neq c_j$ for all i, j , then $\xi_i - \xi_j \neq 0$ and for ε sufficiently small we have $\eta_i - \eta_j$ of opposite sign, thus (12.29) is true.

For a proof of the remaining cases, we shall construct recursively vectors v_0, v_1, v_2, \dots and denote by P_k the partition of $\{1, \dots, s\}$ defined by the equivalence relation

$$i \sim j \iff (v_q)_i = (v_q)_j \quad \text{for } q = 0, 1, \dots, k. \quad (12.30)$$

For a given partition P of $\{1, 2, \dots, s\}$ we introduce the space

$$X(P) = \{v \in \mathbb{R}^s; (v)_i = (v)_j \text{ if } i \sim j \text{ with respect to } P\}.$$

With this notation, the method is S -irreducible if and only if

$$AX(P) \not\subset X(P) \quad (12.31)$$

for every partition other than $\{\{1\}, \{2\}, \dots, \{s\}\}$.

We start with $v_0 = \mathbb{1}$ and $P_0 = \{\{1, \dots, s\}\}$ and define

$$v_{k+1} = \begin{cases} Av_k & \text{if } Av_k \notin X(P_k) \\ \omega & \text{if } Av_k \in X(P_k) \end{cases}$$

where ω is an arbitrary vector of $X(P_k)$ satisfying $A\omega \notin X(P_k)$. Such a choice is possible by (12.31). After a finite number of steps, say m , we arrive at $P_m = \{\{1\}, \{2\}, \dots, \{s\}\}$, because the number of components of P_k is increasing, and strictly increasing after every second step. Therefore all elements of the vector

$$\xi = v_0 - \varepsilon v_1 + \varepsilon^2 v_2 - \dots + (-\varepsilon)^m v_m$$

are distinct (for sufficiently small $\varepsilon > 0$) and (12.29) is satisfied. \square

Lemma 12.20 (Minty 1962). *Let u_1, \dots, u_k and $f(u_1), \dots, f(u_k)$ be elements of \mathbb{R}^n with*

$$\langle f(u_i) - f(u_j), u_i - u_j \rangle < 0 \quad \text{for } i \neq j.$$

Then there exists a continuous extension $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ satisfying

$$\langle f(u) - f(v), u - v \rangle \leq 0 \quad \text{for all } u, v \in \mathbb{R}^n.$$

Proof (Wakker 1985). Define

$$\gamma = \max_{i \neq j} \frac{\langle f(u_i) - f(u_j), u_i - u_j \rangle}{\|f(u_i) - f(u_j)\|^2} < 0$$

and put $g(u_i) = 2\gamma f(u_i) - u_i$, so that $\|g(u_i) - g(u_j)\| \leq \|u_i - u_j\|$. An application of Lemma 12.21 shows that there exists a continuous extension $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ satisfying $\|g(u) - g(v)\| \leq \|u - v\|$ (i.e., g is non-expansive). The function

$$f(u) = \frac{1}{2\gamma}(g(u) + u)$$

then satisfies the requirements. \square

Lemma 12.21 (Kirschbraun 1934). *Let u_1, \dots, u_k and $g(u_1), \dots, g(u_k) \in \mathbb{R}^n$ be such that*

$$\|g(u_i) - g(u_j)\| \leq \|u_i - u_j\| \quad \text{for } i, j = 1, \dots, k. \quad (12.32)$$

Then there exists a continuous extension $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that

$$\|g(u) - g(v)\| \leq \|u - v\| \quad \text{for all } u, v \in \mathbb{R}^n. \quad (12.33)$$

Proof. This was once a difficult result in set-theory. A particularly nice proof, of which we give here a “dynamic” version, has been found by I.J. Schoenberg (1953).

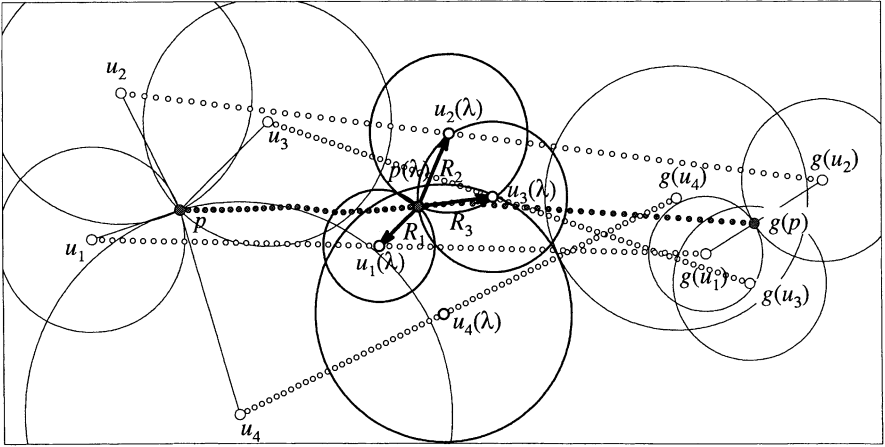


Fig. 12.1. Construction of $g(p)$

a) The main problem is to construct for *one* given point p the extension $g(p)$ such that (12.33) remains satisfied. We move the points u_i into their images $g(u_i)$ by an affine map

$$u_i(\lambda) = u_i + \lambda(g(u_i) - u_i), \quad 0 \leq \lambda \leq 1, \quad i = 1, \dots, k. \quad (12.34)$$

We define $r_i = \|u_i - p\|$ and shrink, for each λ , the balls with center $u_i(\lambda)$ and radius $r_i \mu$ until their intersection consists of one point only

$$\mu(\lambda) := \min \left\{ \mu ; \bigcap_{i=1}^k \{ u ; \|u_i(\lambda) - u\| \leq r_i \mu \} \neq \emptyset \right\}. \quad (12.35)$$

This intersection point, denoted by $p(\lambda)$ (see Fig. 12.1), depends continuously (except for a possible sudden decrease of μ if $\lambda = 0$) and piecewise differentially on λ . We shall show that $\mu(\lambda)$ is non-increasing, which means that $g(p) := p(1)$ is a point satisfying (12.33).

We denote the vectors

$$R_i := u_i(\lambda) - p(\lambda), \quad (12.36)$$

and have from the hypothesis (12.32) that $\|R_i - R_j\|^2$ is non-increasing, hence that $\langle R_i - R_j, dR_i - dR_j \rangle \leq 0$ or

$$\langle R_i, dR_j \rangle + \langle R_j, dR_i \rangle \geq \langle R_i, dR_i \rangle + \langle R_j, dR_j \rangle. \quad (12.37)$$

As can be seen in Fig. 12.1, not all points $u_i(\lambda)$ are always “active” in (12.35), i.e., $p(\lambda)$ lies on the boundary of the shrunk ball centered in $u_i(\lambda)$. While for $\lambda = 0$ (for which $\|R_i\| = r_i\mu$) all four are active, at $\lambda = 1/2$ the active points are $u_1(\lambda)$, $u_2(\lambda)$, $u_3(\lambda)$, and finally for $\lambda = 1$ we only have $u_1(\lambda)$ and $u_2(\lambda)$ active. We suppose, for a given λ , that $u_1(\lambda), \dots, u_m(\lambda)$ ($m \leq k$) are the active points, which may sometimes require a proper renumbering. The crucial idea of Schoenberg is the fact that $p(\lambda)$ lies in the convex hull of $u_1(\lambda), \dots, u_m(\lambda)$, i.e., there are positive values $c_1(\lambda), \dots, c_m(\lambda)$ with $\sum_{i=1}^m c_i R_i = 0$. This means that

$$\langle \sum_i c_i R_i, \sum_j c_j dR_j \rangle = 0.$$

We here apply (12.37) pairwise to i, j and j, i , and obtain

$$0 = \langle \sum_i c_i R_i, \sum_j c_j dR_j \rangle \geq \sum_i \langle R_i, dR_i \rangle (c_i \sum_j c_j).$$

Since by construction (see (12.36)) all $\|R_i\|$ decrease or increase simultaneously with μ , and since all $c_i > 0$, we see that $d\mu \leq 0$, i.e., μ is non-increasing.

b) The rest is now standard (Kirszbraum): we choose a countable dense sequence of points p_1, p_2, p_3, \dots in \mathbb{R}^n and extend g gradually to these points, so that (12.33) is always satisfied. By continuity (see (12.33)), our function is then defined everywhere. This completes the proof of Lemma 12.21 and with it the proof of Theorem 12.18. \square

Nous ne connaissons pas d'exemples de méthodes qui soient B -stables au sens de Butcher et qui ne soient pas B -stables suivant notre définition. (M. Crouzeix 1979)

Remark. Burrage & Butcher (1979) distinguish between BN -stability (based on non-autonomous systems) and B -stability (based on autonomous systems). Since the differential equation constructed in the above proof (see (12.25)) is *autonomous*, both concepts are equivalent for irreducible methods.

Error Growth Function

All the above theory deals only with contractivity when the one-sided Lipschitz constant ν in (12.2) is zero (see Definition 12.2). The question arises whether we can sharpen the estimate when it is known that $\nu < 0$, and whether we can obtain estimates also in the case when (12.2) holds only for some $\nu > 0$.

Definition 12.22 (Burrage & Butcher 1979). Let ν be given and set $x = h\nu$, where h is the step size. We then denote by $\varphi_B(x)$ the smallest number for which the estimate

$$\|y_1 - \hat{y}_1\| \leq \varphi_B(x) \|y_0 - \hat{y}_0\| \quad (12.38)$$

holds for all problems satisfying

$$\operatorname{Re} \langle f(x, y) - f(x, z), y - z \rangle \leq \nu \|y - z\|^2. \quad (12.39)$$

We call $\varphi_B(x)$ the *error growth function* of the method.

We consider here complex-valued functions $f : \mathbb{R} \times \mathbb{C}^n \rightarrow \mathbb{C}^n$. This is not more general (any such system can be written in real form by considering real and imaginary parts, see Eq. (12.4)), but it is more convenient when working with problems $y' = \lambda(x)y$, where $\lambda(x)$ is complex-valued.

In the case of a linear nonautonomous problem $y' = A(x)y$, condition (12.39) becomes $\mu(A(x)) \leq \nu$ (where $\mu(\cdot)$ denotes the logarithmic norm; see Sect. I.10). Putting $Z_i := hA(x_0 + c_i h)$, the difference of two numerical solutions becomes

$$y_1 - \hat{y}_1 = K(Z_1, \dots, Z_s)(y_0 - \hat{y}_0),$$

where

$$K(Z_1, \dots, Z_s) = I + (b^T \otimes I)Z(I \otimes I - (A \otimes I)Z)^{-1}(\mathbb{1} \otimes I), \quad (12.40)$$

and Z is the block diagonal matrix with Z_1, \dots, Z_s as entries in the diagonal.

Theorem 12.23. *The error growth function of an implicit Runge-Kutta method satisfies*

$$\varphi_B(x) = \sup_{\mu(Z_1) \leq x, \dots, \mu(Z_s) \leq x} \|K(Z_1, \dots, Z_s)\|. \quad (12.41)$$

Proof. Upper Bound. The difference $\Delta y_1 = y_1 - \hat{y}_1$ of two Runge-Kutta solutions satisfies (12.5). The assumption (12.39) implies that $\operatorname{Re} \langle \Delta f_i, \Delta g_i \rangle \leq x \|\Delta g_i\|^2$. We shall prove that there exist matrices Z_i ($i = 1, \dots, s$) with $\mu(Z_i) \leq x$ such that $\Delta f_i = Z_i \Delta g_i$. This implies $\Delta y_1 = K(Z_1, \dots, Z_s) \Delta y_0$ and, as a consequence, that the right-hand expression of Eq. (12.41) is an upper bound of $\varphi_B(x)$.

If $\Delta g_i = 0$ then $\Delta f_i = 0$ and we can take an arbitrary matrix satisfying $\mu(Z_i) \leq x$. Therefore, let us consider vectors f, g (with $g \neq 0$) in \mathbb{C}^n satisfying $\operatorname{Re} \langle f, g \rangle \leq x \|g\|^2$. We put $u_1 := g/\|g\|$, and complete it to an orthonormal basis u_1, \dots, u_n

of \mathbb{C}^n . Then we define the matrix Z by

$$Zu_1 := f/\|g\|, \quad Zu_i := xu_i - \langle u_i, f \rangle u_1 / \|g\|, \quad i = 2, \dots, n.$$

We have $Zg = f$, and one readily verifies that $\operatorname{Re} \langle Zv, v \rangle \leq x\|v\|^2$ for all $v = \sum_{i=1}^n \alpha_i u_i$.

Lower Bound. We first consider nonconfluent Runge-Kutta methods. For given Z_1, \dots, Z_s with $\mu(Z_i) \leq x$ let $A(x)$ be a continuous function satisfying $hA(x_0 + c_i h) = Z_i$ and $\mu(A(x)) \leq x$ for all x ($A(x)$ is, for example, obtained by linear interpolation). Then we have $\Delta y_1 = K(Z_1, \dots, Z_s) \Delta y_0$ and, consequently, also $\varphi_B(x) \geq \|K(Z_1, \dots, Z_s)\|$ for all Z_1, \dots, Z_s with $\mu(Z_i) \leq x$.

For confluent methods the proof is more complicated. Without loss of generality we can assume that the method is irreducible, because neither the value $\varphi_B(x)$ nor the right-hand expression of Eq. (12.41) change, when the method is replaced by an equivalent one. The main observation is now that the Lemmata 12.20 and 12.21 are valid in arbitrary dimensions. Consider Z_1, \dots, Z_s with $\mu(Z_i) \leq x$, such that the linear system $\Delta g_i = \Delta y_0 + \sum_{j=1}^s a_{ij} Z_j \Delta g_j$ has a solution. Exactly as in the proof of Theorem 12.18 we can construct a continuous function $f: \mathbb{C}^n \rightarrow \mathbb{C}^n$, which satisfies (12.39) with $\nu = x$ (we put $h = 1$) and $f(g_i) - f(\hat{g}_i) = Z_i(g_i - \hat{g}_i)$. This completes the proof of the theorem. \square

For 1-stage methods ($s = 1$) the Theorem of von Neumann (Corollary 11.4) implies that it is sufficient to consider scalar, complex-valued z_1 in Eq. (12.41). Since $K(z) = R(z)$ in this case, we have

$$\varphi_B(x) = \varphi_R(x) \quad \text{for all 1-stage methods.} \quad (12.42)$$

For the moment it is not clear, whether one can restrict the supremum in Eq. (12.41) to scalar, complex-valued z_i also for $s \geq 2$. This would require a generalization of the Theorem of von Neumann to functions of more than one variables (Hairer & Wanner 1996). We shall come back to this question later in this section.

Theorem 12.24 (Hairer & Zennaro 1996). *For B -stable Runge-Kutta methods the error growth function is superexponential, i.e., $\varphi_B(0) = 1$ and*

$$\varphi_B(x_1) \varphi_B(x_2) \leq \varphi_B(x_1 + x_2) \quad \text{for } x_1, x_2 \text{ having the same sign.}$$

Proof. The property $\varphi_B(0) = 1$ follows from Definition 12.5. For the proof of the inequality we consider the rational function

$$S(z) = u_A^* K(A_1 - zI, \dots, A_s - zI) v_A u_B^* K(B_1 + zI, \dots, B_s + zI) v_B,$$

where the matrices A_j, B_j satisfy $\mu(A_j) \leq x_1 + x_2$ and $\mu(B_j) \leq 0$, and u_A, v_A, u_B, v_B are arbitrary vectors of \mathbb{C}^n . Using the property $\mu(A_j - zI) = \mu(A_j) - \operatorname{Re} z$ and the fact that $\|C\| = \sup_{\|u\|=1, \|v\|=1} |u^* C v|$, the inequality is obtained exactly as in the proof of Theorem 11.6. \square

The fact that $\varphi_B(x)$ is superexponential together with $\varphi_B(-\infty) = |R(\infty)|$ (see Exercise 8) allows us to draw the same conclusions on asymptotic stability of numerical solutions as in Sect. IV.11.

Computation of $\varphi_B(x)$

The idea is to search for the maximum of $\|\Delta y_1\|$ under the restriction (12.39). More precisely, we consider the following inequality constrained optimization problem:

$$\begin{aligned} \|\Delta y_1\|^2 &\rightarrow \max, \\ \operatorname{Re} \langle \Delta f_i, \Delta g_i \rangle &\leq x \|\Delta g_i\|^2, \quad i = 1, \dots, s. \end{aligned} \quad (12.43)$$

Here $\Delta f_1, \dots, \Delta f_s$ are regarded as independent variables in \mathbb{C}^n , Δy_1 and Δg_i are defined by (12.5), and Δy_0 is considered as a parameter. A classical approach for solving the optimization problem (12.43) is to introduce Lagrange multipliers d_1, \dots, d_s , and to consider the Lagrangian

$$\begin{aligned} \mathcal{L}(\Delta f, D) &= \frac{1}{2} \|\Delta y_1\|^2 - \sum_{i=1}^s d_i \left(\operatorname{Re} \langle \Delta f_i, \Delta g_i \rangle - x \|\Delta g_i\|^2 \right) \\ &= -\frac{1}{2} (\Delta y_0^*, \Delta f^*) \left(\begin{pmatrix} \alpha & u^T \\ u & W \end{pmatrix} \otimes I \right) \begin{pmatrix} \Delta y_0 \\ \Delta f \end{pmatrix}, \end{aligned} \quad (12.44)$$

where $\Delta f = (\Delta f_1, \dots, \Delta f_s)^T$, $D = \operatorname{diag}(d_1, \dots, d_s)$, and

$$\alpha = -1 - 2x \mathbb{1}^T D \mathbb{1}, \quad (12.45a)$$

$$u = D \mathbb{1} - b - 2x A^T D \mathbb{1}, \quad (12.45b)$$

$$W = DA + A^T D - bb^T - 2x A^T D A. \quad (12.45c)$$

Theorem 12.25 (Burrage & Butcher 1980). *If the matrix*

$$\begin{pmatrix} \alpha + \varphi^2 & u^T \\ u & W \end{pmatrix} \quad \text{is positive semi-definite} \quad (12.46)$$

for some $d_1 \geq 0, \dots, d_s \geq 0$, then it holds $\|\Delta y_1\| \leq \varphi \|\Delta y_0\|$ for all problems satisfying (12.39) with $h\nu \leq x$. Consequently, we have $\varphi_B(x) \leq \varphi$.

Proof. Subtracting $\varphi^2 \|\Delta y_0\|^2 / 2$ from both sides of (12.44) yields

$$\frac{1}{2} \left(\|\Delta y_1\|^2 - \varphi^2 \|\Delta y_0\|^2 \right) - \sum_{i=1}^s d_i \left(\operatorname{Re} \langle \Delta f_i, \Delta g_i \rangle - x \|\Delta g_i\|^2 \right) \leq 0.$$

The statement then follows from $d_i \geq 0$ and $\operatorname{Re} \langle \Delta f_i, \Delta g_i \rangle \leq x \|\Delta g_i\|^2$. \square

With the help of this theorem, Burrage & Butcher (1980) computed an upper bound of $\varphi_B(x)$ for many 2-stage methods. It turned out that for all these 2-stage methods $\varphi_B(x) = \varphi_K(x)$, where

$$\varphi_K(x) = \sup_{\operatorname{Re} z_1 \leq x, \dots, \operatorname{Re} z_s \leq x} |K(z_1, \dots, z_s)|. \quad (12.47)$$

There naturally arises the question: *is it true that $\varphi_B(x) = \varphi_K(x)$ for all Runge-Kutta methods?* If we want to check the validity of $\varphi_B(x) = \varphi_K(x)$ for a given Runge-Kutta method, we have to find non-negative Lagrange multipliers d_i , such that (12.46) is satisfied. The following lemmas will be useful for this purpose.

We denote by z_1^0, \dots, z_s^0 the values, for which the supremum in Eq. (12.47) is attained. By the maximum principle we have $z_j^0 = x + iy_j^0$ ($y_j^0 = \infty$ is admitted). We further put $z^0 = (z_1^0, \dots, z_s^0)$ and let $\partial_j K(z^0)$ be the derivative of $K(z_1, \dots, z_s)$ with respect to the j th argument, evaluated at z^0 .

Lemma 12.26. *Let x be fixed with $\varphi_K(x) < \infty$. The condition (12.46) with $\varphi = \varphi_K(x)$ then uniquely determines the Lagrange multipliers d_1, \dots, d_s (see Eq. (12.53) below). They are real and positive.*

Proof. Consider the identity (12.44) for the special case, where Δf_j is scalar, $\Delta f_j = z_j \Delta g_j$, and hence $\Delta y_1 = K(z_1, \dots, z_s)$. For $\operatorname{Re} z_j = x$ this identity becomes

$$|K(z_1, \dots, z_s)|^2 - \varphi^2 = -(1, \Delta f^*) \begin{pmatrix} \alpha + \varphi^2 & u^T \\ u & W \end{pmatrix} \begin{pmatrix} 1 \\ \Delta f \end{pmatrix}. \quad (12.48)$$

Putting $\varphi := \varphi_K(x)$ and $z_j := z_j^0$ (eventually one has to consider limits) the left-hand expression of Eq. (12.48) vanishes. This together with assumption (12.46) implies that $u + W \Delta f = 0$, i.e.,

$$D \mathbb{1} - b - 2x A^T D \mathbb{1} + (DA + A^T D - bb^T - 2x A^T DA) \Delta f = 0.$$

Collecting suitable terms, and using $\Delta f = Z_0 \Delta g$ and $\Delta g = \mathbb{1} + A \Delta f$, where $Z_0 = \operatorname{diag}(z_1^0, \dots, z_s^0)$, this relation becomes

$$D \Delta g = (I - A^T Z_0^*)^{-1} b \cdot K(z^0). \quad (12.49)$$

We shall show that all components of $\Delta g = (I - AZ_0)^{-1} \mathbb{1}$ are different from zero, so that (12.49) uniquely determines the Lagrange multipliers d_1, \dots, d_s .

Expanding $K(z_1, \dots, z_s)$ into a Taylor series with respect to z_j , we obtain

$$K(z_1^0, \dots, z_j, \dots, z_s^0) = K(z^0) \left(1 + c(z_j - z_j^0) + \mathcal{O}((z_j - z_j^0)^2) \right),$$

where $c = \partial_j K(z^0)/K(z^0)$. Since $|K(z_1^0, \dots, z_j, \dots, z_s^0)| \leq |K(z^0)|$ for $\operatorname{Re} z_j \leq \operatorname{Re} z_j^0$, we have $c > 0$, and consequently also

$$\partial_j K(z^0) \neq 0, \quad 0 < \partial_j K(z^0)/K(z^0) < \infty. \quad (12.50)$$

Differentiating $K(z_1, \dots, z_s) = 1 + b^T Z(I - AZ)^{-1} \mathbb{1}$ with respect to z_j yields

$$\partial_j K(z^0) = b^T (I - Z_0 A)^{-1} e_j e_j^T (I - AZ_0)^{-1} \mathbb{1}, \quad (12.51)$$

and we obtain from (12.50) that

$$b^T (I - Z_0 A)^{-1} e_j \neq 0, \quad \Delta g_j = e_j^T (I - AZ_0)^{-1} \mathbb{1} \neq 0, \quad (12.52)$$

so that d_1, \dots, d_s are uniquely determined by (12.49). Dividing the j th component of (12.49) by Δg_j , it follows from (12.51) that

$$d_j = |b^T (I - Z_0 A)^{-1} e_j|^2 \cdot \frac{K(z^0)}{\partial_j K(z^0)}, \quad (12.53)$$

which is a strictly positive real number by (12.50) and (12.52).

In this proof we have implicitly assumed that all z_j^0 are finite. If $z_j^0 = x + i\infty$ for some j , one has to apply the standard transformation $\omega_j = x + 1/(z_j - x)$, which maps the half-plane $\operatorname{Re} z_j \leq x$ onto $\operatorname{Re} \omega_j \leq x$, and ∞ into 0. \square

Lemma 12.27. *If the matrix W of Eq. (12.45c), with d_1, \dots, d_s given by Lemma 12.26, is positive semi-definite, then we have $\varphi_B(x) = \varphi_K(x)$.*

Proof. It follows from

$$\begin{pmatrix} \alpha + \varphi_K^2(x) & u^T \\ u & W \end{pmatrix} \begin{pmatrix} 1 \\ \Delta f \end{pmatrix} = 0 \quad (12.54)$$

(see Eq. (12.48)) and from $v^T W v \geq 0$ for all $v \in \mathbb{R}^s$ that the matrix in (12.54) is positive semi-definite. The statement then follows from Theorem 12.25. \square

With the above results it is possible to check for a given Runge-Kutta method, whether $\varphi_B(x) = \varphi_K(x)$ is satisfied. This can be done by the following algorithm:

- compute $\varphi = \varphi_K(x)$ of Eq. (12.47) either numerically or with help of a formula manipulation program;
- compute the Lagrange multipliers d_1, \dots, d_s from Lemma 12.26;
- check, whether the matrix W of Eq. (12.45c) is positive semi-definite. If this is the case, it holds $\varphi_B(x) = \varphi_K(x)$ by Lemma 12.27.

Example 12.28. For the two-stage Radau IIA method (see Table 5.5) the function $K(z_1, z_2)$ is given by

$$K(z_1, z_2) = \frac{1 + z_1/3}{1 - 5z_1/12 - z_2/4 + z_1 z_2/6}.$$

The maximum of $|K(z_1, z_2)|$ on the set $\operatorname{Re} z_i \leq x$ is attained at

$$z_1^0 = \begin{cases} x + i\infty & \text{for } x \leq \xi \\ x + ix\sqrt{\frac{45 - 42x + 8x^2}{9 + 18x - 8x^2}} & \text{for } \xi \leq x < 3/2 \end{cases}$$

$$z_2^0 = \begin{cases} x & \text{for } x \leq \xi \\ x + i\frac{x\sqrt{(45 - 42x + 8x^2)(9 + 18x - 8x^2)}}{8x^2 - 6x - 9} & \text{for } \xi \leq x < 3/2 \end{cases}$$

(the value $\xi = (9 - 3\sqrt{17})/8$ is a root of $9 + 18x - 8x^2 = 0$) and it is given by

$$\varphi_K(x) = \begin{cases} \frac{4}{5 - 2x} & \text{if } x \leq \xi \\ \frac{3 + 4x}{\sqrt{(3 - 2x)(3 + 4x - 2x^2)}} & \text{if } \xi \leq x < 3/2. \end{cases}$$

The function $K(z_1, z_2)$ is not bounded on $\operatorname{Re} z_i \leq x$ for $x \geq 3/2$. From the proof of Lemma 12.26 we compute d_1 and d_2 , and obtain

$$d_1 = \begin{cases} \frac{9}{(3 - x)(5 - 2x)} & \text{for } x \leq \xi \\ \frac{(3 + 4x)^2}{4(3 + 4x - 2x^2)} & \text{for } \xi \leq x \end{cases} \quad d_2 = \begin{cases} \frac{2}{5 - 2x} & \text{for } x \leq \xi \\ \frac{3 + 4x}{4(3 + 4x - 2x^2)} & \text{for } \xi \leq x. \end{cases}$$

With these values one checks straight-forwardly that the matrix W of Eq. (12.45c) is semi-definite positive, so that $\varphi_B(x) = \varphi_K(x)$; see also Burrage & Butcher (1980). Actually, the matrix W is non-singular for $x < \xi$, and of rank one for $\xi \leq x < 3/2$.

A comparison with Eq. (11.15) shows that we do not obtain the same estimate as for linear autonomous problems.

The above algorithm can easily be applied to other two-stage methods. We thus obtain for the two-stage Gauss method

$$\varphi_B(x) = \begin{cases} 1 & \text{if } -\infty < x \leq 0 \\ \frac{2x + \sqrt{9 + 3x^2}}{3 - x} & \text{if } 0 \leq x < 3, \end{cases}$$

and for the two-stage Lobatto IIIC method

$$\varphi_B(x) = \begin{cases} \frac{1}{1 - x + x^2} & \text{if } -\infty < x \leq 0 \\ \frac{1}{1 - x} & \text{if } 0 \leq x < 1. \end{cases}$$

For methods with more than two stages, explicit formulas are difficult to obtain, and one has to apply numerical methods for the computation of z_j^0 (supremum in Eq. (12.47)).

Exercises

1. Prove, directly from Def. 12.2, that the implicit Euler method is B -stable.
2. Let M be a symmetric $s \times s$ -matrix and $\langle \cdot, \cdot \rangle$ the scalar product of \mathbb{R}^n . Then M is non-negative definite, if and only if

$$\sum_{i=1}^s \sum_{j=1}^s m_{ij} \langle u_i, u_j \rangle \geq 0 \quad \text{for all } u_i \in \mathbb{R}^n.$$

Hint. Use $M = Q^T D Q$ where D is diagonal.

3. Give a simple proof for the B -stability of the Radau IIA methods by extending the ideas of Example 12.3.

Hint. For the quadrature, based on the zeros of (5.2), we have

$$\int_0^1 \varphi(x) dx = \sum_{i=1}^s b_i \varphi(c_i) + C \varphi^{(2s-1)}(\xi), \quad 0 < \xi < 1.$$

with $C < 0$ (see e.g. Abramowitz & Stegun (1964, Formula 25.4.31)).

4. (Dahlquist & Jeltsch 1987). Prove that Method I of Table 12.4 is S -reducible with respect to the partition $(\{1\}, \{2, 3\})$. The reduced method II itself is DJ -reducible and reduces to Method III.

For the initial value problem $y' = f(y)$, $y(0) = 1$, where $f(y) = y^2$ for $y \geq 0$ and $f(y) = 0$ for $y < 0$, and for $h = 2$, Methods I and III have unique solutions which are different. Explain this apparent contradiction.

Table 12.4. Reduction of RK-methods

0	0	0	0	0	0	0
1/2	0	1	-1/2	0	0	0
1/2	0	1/2	0	0	1/2	0
	1	b	$-b$	1	0	1
Method I				Method II		Method III

5. Give a counterexample of an irreducible AN -stable but not algebraically stable, and hence not B -stable method.

Hint. Start with any algebraically stable method with, say, two stages and modify it as indicated in Table 12.5. Find conditions on the free parameters d, e, α such that the two methods are identical for equations $y' = \lambda(x)y$. This ensures AN -stability of the second method. Then play with the parameters to destroy algebraic stability.

6. Show that the method of Table 12.1 is DJ -reducible, but not S -reducible; show that it is algebraically stable together with the reduced method.

Table 12.5. Construction of AN -stable but not B -stable method

c_1	a_{11}	a_{12}	\Rightarrow	c_1	a_{11}	$a_{12}\alpha$	$a_{12}(1-\alpha)$
c_2	a_{21}	a_{22}		c_2	$c_2 - d$	$d\alpha$	$d(1-\alpha)$
				c_2	$c_2 - e$	$e\alpha$	$e(1-\alpha)$
	b_1	b_2			b_1	$b_2\alpha$	$b_2(1-\alpha)$

Show that the method of Table 12.2 is S -reducible, but not DJ -reducible; show that it is not algebraically stable, but that the reduced method is.

7. (Sandberg & Shichman 1968, Vanselow 1979, Hundsdorfer 1985). Prove that Rosenbrock methods are not B -stable in the sense of Definition 11.2.

Hint. Apply the method to the scalar problem $y' = f(y)$, $y_0 = 1$ where $f(y)$ is a non-increasing function satisfying (for a small ε)

$$f(y) = \begin{cases} -y & \text{if } |y-1| \geq 2\varepsilon \\ -1 & \text{if } |y-1| \leq \varepsilon. \end{cases}$$

8. (Hairer & Zennaro 1996). For irreducible, algebraically stable Runge-Kutta methods the error growth function satisfies

$$\varphi_B(x) \leq \frac{\sqrt{1-2x\gamma(1-\varrho^2)} - 2x\gamma\varrho}{1-2x\gamma} \quad \text{for } x \leq 0,$$

where $\varrho = |R(\infty)|$ ($R(z)$ is the stability function), $\gamma = \left(\sum_{j=1}^s b_j^{-1} v_j^2\right)^{-1}$, and $(v_1, \dots, v_s)^T = \lim_{\varepsilon \rightarrow 0} b^T(A + \varepsilon I)^{-1}$.

Hint. From (12.7) we have $\|\Delta y_1\|^2 \leq \|\Delta y_0\|^2 + 2x \sum_i b_i \|\Delta g_i\|^2$. Then, compute Δf_i from (12.5b) (if A is invertible), insert it into (12.5a) and conclude $\Delta y_1 = R(\infty)\Delta y_0 + \sum_j (\sum_i b_i \omega_{ij}) \Delta g_j$, where $(\omega_{ij}) = A^{-1}$. The Cauchy-Schwarz inequality yields $\sum_i b_i \|\Delta g_i\|^2 \geq \gamma (\|\Delta y_1\| - \varrho \|\Delta y_0\|)^2$ which, inserted into the first estimate, gives a second degree inequality for Δy_1 .

9. Prove that for the 3-stage Gauss method we have for $x \geq 0$

$$\varphi_B(x) \geq (1+x/2)/(1-x/2).$$

Hint. Using (12.18), compute $K(Z)$ for $z_1 \rightarrow -\infty$, $z_2 = x$, $z_3 \rightarrow -\infty$.

10. If the matrix W of Eq. (12.45c), with d_1, \dots, d_s given by Lemma 12.26, is either non-singular or of rank ≤ 1 , then it holds $\varphi_B(x) = \varphi_K(x)$.

Hint. Exploit the fact that the expression in Eq. (12.48) with $\varphi = \varphi_K(x)$ is non-positive for all z_j with $\operatorname{Re} z_j \leq x$.

11. Show that for a non-negative definite symmetric matrix $M = (m_{ij})$ one has

$$|m_{ij}| \leq \sqrt{m_{ii}m_{jj}}.$$

IV.13 Positive Quadrature Formulas and B-Stable RK-Methods

Bien que le problème (des quadratures) ait une durée de deux cents ans à peu près, bien qu'il était l'objet de nombreuses recherches de plusieurs géomètres: Newton, Cotes, Gauss, Jacobi, Hermite, Tchébychef, Christoffel, Heine, Radeau [sic], A. Markov, T. Stijtes [sic], C. Possé, C. Andreev, N. Sonin et d'autres, il ne peut être considéré, cependant, comme suffisamment épuisé.

(V. Steklov 1917)

We shall give a constructive characterization of all irreducible B -stable Runge-Kutta methods (Theorem 13.15). Because of Theorem 12.16 we first have to study quadrature formulas with positive weights.

Quadrature Formulas and Related Continued Fractions

Steklov (1916) proved that a family of interpolatory quadrature formulas converges for all Riemann integrable functions, if all weights of the formulas are positive ("Il faut remarquer cependant que de tels théorèmes généraux ne peuvent avoir aucune valeur pratique ..."). This theorem, rediscovered around 1922 by Fejér, initiated an extensive search for quadrature formulas with positive weights. Fejér (1933, "weiter habe ich noch auf sehr kurzem Wege das folgende Resultat erhalten ...") found the result:

"If $P_s(z)$ are the Legendre polynomials normalized as in (13.4) and c_1, \dots, c_s are the zeros of $M(z) = P_s(z) + \alpha_1 P_{s-1}(z) + \alpha_2 P_{s-2}(z)$ with $\alpha_2 \leq 0$, then the weights b_i are all positive".

The theory of B -stable methods renewed the interest in positive quadrature formulas and Burrage (1978) obtained the sharp bound

$$\alpha_2 < \frac{(s-1)^2}{4(2s-1)(2s-3)} \quad (13.1)$$

for the positivity of the b_i in the above case. This is the same as condition (5.51) in a different normalization. A short proof of this result (see "Lemma 18" of Nørsett & Wanner 1981) then led to a complete characterization of positive quadrature formulas by Sottas & Wanner (1982). An independent proof of an equivalent result was found by Peherstorfer (1981). In what follows, we give a new approach using continued fractions.

Consider a quadrature formula

$$\sum_{j=1}^s b_j f(c_j) \approx \int_0^1 f(x) dx$$

with distinct nodes c_i and non-zero weights b_i . The main idea is to consider the rational function

$$Q(z) = \sum_{j=1}^s b_j \frac{1}{z - c_j} = \frac{N(z)}{M(z)} \quad (13.2)$$

where, as usual, $M(z) = (z - c_1) \cdot \dots \cdot (z - c_s)$. We first express the order of the quadrature formula in terms of the function $Q(z)$.

Lemma 13.1. *A quadrature formula is of order p if and only if $Q(z)$, defined by (13.2), satisfies*

$$Q(z) = -\log\left(1 - \frac{1}{z}\right) + \mathcal{O}\left(\frac{1}{z^{p+1}}\right) \quad \text{for } z \rightarrow \infty. \quad (13.3)$$

Proof. Inserting the geometric series for $(1 - c_j/z)^{-1}$ into (13.2) we obtain

$$Q(z) = \sum_{k \geq 1} \left(\sum_{j=1}^s b_j c_j^{k-1} \right) \frac{1}{z^k}.$$

Therefore (13.3) is equivalent to

$$\sum_{j=1}^s b_j c_j^{k-1} = \frac{1}{k} \quad \text{for } k = 1, \dots, p. \quad \square$$

We now study the case of the *Gaussian quadrature formulas*, where the function (13.2) will be denoted by $Q_s^G(z) = N_s^G(z)/M_s^G(z)$; here the c_i are the zeros of the s -degree shifted Legendre polynomial

$$P_s(z) = \frac{s!}{(2s)!} \frac{d^s}{dz^s} (z^s(z-1)^s), \quad (13.4)$$

which is normalized so that the coefficient of z^s is 1. The polynomials (13.4) satisfy the recurrence relation (see Eq. (5.53) or Abramowitz & Stegun, p. 782)

$$P_{s+1}(z) = \left(z - \frac{1}{2}\right) P_s(z) - \tau_s P_{s-1}(z), \quad \tau_s = \frac{s^2}{4(4s^2 - 1)} \quad (13.5)$$

and $P_0(z) = 1$, $P_{-1}(z) = 0$. Since this quadrature formula is of optimal order $2s$, it follows from (13.3) that

$$N_s^G(z) = -M_s^G(z) \log\left(1 - \frac{1}{z}\right) + \mathcal{O}\left(\frac{1}{z^{s+1}}\right). \quad (13.6)$$

We now insert $M_s^G(z) = P_s(z)$ (see (13.2)) into (13.5) and multiply by $\log(1 - 1/z)$ (which is $\mathcal{O}(1/z)$ for $z \rightarrow \infty$). A comparison with (13.6) shows that the polynomials $N_s^G(z)$ must also satisfy the recurrence formula (13.5) (with $N_0^G(z) = 0$, $N_1^G(z) = 1$). It thus follows from elementary properties of continued fractions

(Exercise 1 or Perron (1913), page 4) that

$$Q_s^G(z) = \frac{1}{z - \frac{1}{2}} - \frac{\tau_1}{z - \frac{1}{2}} - \cdots - \frac{\tau_{s-1}}{z - \frac{1}{2}}. \quad (13.7)$$

For an arbitrary quadrature formula we have

Lemma 13.2. *An irreducible rational function $Q(z) = N(z)/M(z)$ (with $\deg M = s$, $\deg N = s - 1$) satisfies (13.3) with $p \geq 2(s - k)$, if and only if*

$$Q(z) = \frac{1}{z - \frac{1}{2}} - \frac{\tau_1}{z - \frac{1}{2}} - \cdots - \frac{\tau_{s-k-1}}{z - \frac{1}{2}} - \frac{g(z)}{f(z)} \quad (13.7')$$

with $\deg f = k$ and $\deg g \leq k - 1$.

Proof. From Lemma 13.1 we know that $Q(z) = Q_s^G(z) + \mathcal{O}(1/z^{2(s-k)+1})$. Therefore the first $2(s - k)$ coefficients in the continued fraction expansions for $Q(z)$ and $Q_s^G(z)$ must be the same. \square

Endlich sei noch die folgende Formel wegen ihrer häufigen Anwendungen ausdrücklich hervorgehoben:

(O. Perron 1913, page 5)

Lemma 13.3. *The functions $M(z)$ and $N(z)$ of Lemma 13.2 are related to $f(z)$ and $g(z)$ of (13.7') as follows:*

$$\begin{aligned} M(z) &= P_{s-k}(z)f(z) - P_{s-k-1}(z)g(z), \\ N(z) &= N_{s-k}^G(z)f(z) - N_{s-k-1}^G(z)g(z). \end{aligned} \quad (13.8)$$

Proof. This follows from the recursion (13.30) and Exercise 1 below, if we put there $b_0 = 0$, $b_1 = \cdots = b_{s-k} = z - 1/2$, $b_{s-k+1} = f(z)$ and $a_1 = 1$, $a_j = -\tau_{j-1}$ ($j = 2, \dots, s - k$), $a_{s-k+1} = -g(z)$. \square

Solving the linear system (13.8) for $f(z)$ and $g(z)$ gives, with the use of Exercise 2,

$$\begin{aligned} f(z) \cdot \tau_1 \cdots \tau_{s-k-1} &= N(z)P_{s-k-1}(z) - M(z)N_{s-k-1}^G(z) \\ g(z) \cdot \tau_1 \cdots \tau_{s-k-1} &= N(z)P_{s-k}(z) - M(z)N_{s-k}^G(z). \end{aligned} \quad (13.9)$$

Number of Positive Weights

For a given rational function (13.2), the weights are determined by

$$b_i = \frac{N(c_i)}{M'(c_i)}. \quad (13.10)$$

But we want our theory to work also for *confluent* nodes for which $M'(c_i) = 0$.

Therefore we suppose that c_1, \dots, c_m ($m \leq s$) are the *real and distinct* zeros of $M(z)$ of *multiplicities* l_1, \dots, l_m . Then we let

$$b_i = \frac{N(c_i)}{M^{(l_i)}(c_i)} \quad i = 1, \dots, m. \quad (13.10')$$

For $l_i = 1$ this is just (13.10); otherwise we are considering the weights for the highest derivative of a Hermitian quadrature formula (see Exercise 3).

The main idea (following Sottas & Wanner 1982) is now to consider the path $\gamma(t) = (f(t), g(t))$ in the plane \mathbb{R}^2 , where f and g are the polynomials of (13.7'). For $t \rightarrow \pm\infty$ this path tends to infinity with horizontal limiting directions, since the degree of f is higher than that of g . Equation (13.8) tells us that for an irreducible $Q(z)$ this path does not pass through the origin.

Definition 13.4. The *rotation number* r of γ is the integer for which $r\pi$ is the total angle of rotation around the origin for the path $\gamma(t)$ ($-\infty < t < \infty$) measured in the negative (clockwise) sense. Counter-clockwise rotations are negative.

An algebraic definition of r is possible as

$$r = \sum_i \text{sign} (f^{(l_i)}(t_i)g(t_i)),$$

where the summation is over all real zeros of $f(t)$ with *odd* multiplicity l_i .

Theorem 13.5 (Sottas & Wanner 1982). *Let $Q(z) = N(z)/M(z)$ be an irreducible rational function as in Lemma 13.2. Suppose that c_1, \dots, c_m are the (distinct) real zeros of $M(z)$ with odd multiplicity and denote by n_+ (respectively n_-) the number of positive (respectively negative) b_i . Further, let r be the rotation number of $\gamma = (f, g)$ (Definition 13.4). Then*

$$n_+ - n_- = s - k + r. \quad (13.11)$$

Proof. The proof is by counting the number of crossings of the vectors $\gamma(t) = (f(t), g(t))$ and $\beta(t) = (P_{s-k-1}(t), P_{s-k}(t))$, like the crossings of hands on a Swiss cuckoo clock.

From (13.9) we see that when t equals a zero c_i of M , these two vectors are parallel in the same sense ($N(c_i) > 0$) or in the opposite sense ($N(c_i) < 0$). From (13.8) we observe that $M(t)$ is just the exterior product $\gamma(t) \times \beta(t)$. By elementary geometry, and taking into account Formula (13.10'), we see that at every zero c_i with odd multiplicity we have

- i) $b_i > 0$, if the crossing of $\gamma(t)$ with $\beta(t)$ is clockwise;
- ii) $b_i < 0$, if this crossing is counter-clockwise.

Zeros of $M(t)$ with even multiplicity don't give rise to crossings.

Since the zeros of P_{s-k} and P_{s-k-1} interlace (see e.g. Theorem 3.3.2 of Szegő 1939), the vector $\beta(t)$ turns counter-clockwise with a total angle of $-(s -$

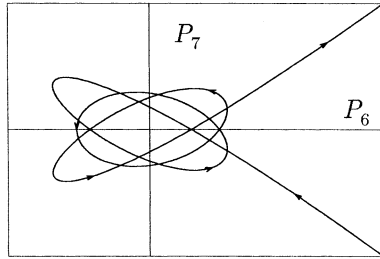


Fig. 13.1. The path $(P_{s-k-1}(t), P_{s-k}(t))$ for $s - k = 7$

$k)\pi$ (see Fig. 13.1). The vector $\gamma(t)$ turns with a total angle $r\pi$ measured clockwise (Definition 13.4). Since the limiting directions of $\gamma(t)$ and $\beta(t)$ are different (horizontal for $\gamma(t)$ and vertical for $\beta(t)$), $\gamma(t)$ must cross $\beta(t)$, as t increases from $-\infty$ to $+\infty$, exactly $s - k + r$ times more often clockwise than counter-clockwise. This gives Formula (13.11). \square

Corollary 13.6. *Under the assumptions of Theorem 13.5, all zeros of $M(z)$ are real and simple, and the b_i are positive if and only if*

$$r = k.$$

Proof. $r = k$ means by (13.11) that $n_+ - n_- = s$. Because of $n_- \geq 0$ and $n_+ \leq s$, this is equivalent to $n_+ = s$ and $n_- = 0$. \square

Characterization of Positive Quadrature Formulas

The following theorem gives a constructive characterization of all quadrature formulas with positive weights.

Theorem 13.7. *Let*

$$\sigma_1 < \varrho_1 < \sigma_2 < \varrho_2 < \dots < \varrho_{k-1} < \sigma_k$$

be arbitrary real numbers and C a positive constant. Then, putting

$$f(z) = (z - \sigma_1) \dots (z - \sigma_k), \quad g(z) = C(z - \varrho_1) \dots (z - \varrho_{k-1}), \quad (13.12)$$

computing $M(z)$, $N(z)$ from (13.8), taking c_1, \dots, c_s as the zeros of $M(z)$ and b_i from (13.10), one obtains all quadrature formulas with positive weights of order $p \geq 2(s - k)$. If $C = \tau_{s-k}$ the order is $p \geq 2(s - k) + 1$.

Proof. The functions $f(z)$ and $g(z)$ are irreducible, so that also the fraction $N(z)/M(z)$ is irreducible by (13.9). The statement now follows from Corollary 13.6, since the polynomials (13.12) are all possible polynomials for which $r = k$. The stated order properties follow from Lemma 13.2. \square

Example 13.8. Let c_1, \dots, c_s be the zeros of

$$M(z) = P_s(z) + \alpha_1 P_{s-1}(z) + \alpha_2 P_{s-2}(z). \quad (13.13)$$

In order to study when the corresponding quadrature formula has positive weights, we use (13.5) to write (13.13) as

$$M(z) = P_{s-1}(z) \left(z - \frac{1}{2} + \alpha_1 \right) - P_{s-2}(z) (\tau_{s-1} - \alpha_2).$$

Consequently $f(z) = z - 1/2 + \alpha_1$, $g(z) = \tau_{s-1} - \alpha_2$ and Theorem 13.7 implies that the zeros of $M(z)$ are real and the weights positive, if and only if $\alpha_2 < \tau_{s-1}$, hence (13.1) is proved.

For $k > 1$ the rotation number r of $(f(t), g(t))$ can be computed with Sturm's algorithm (Lemma 13.3 of Sect. I.13). Consider, for example,

$$\begin{aligned} M(z) &= P_s(z) + \alpha_1 P_{s-1}(z) + \alpha_2 P_{s-2}(z) + \alpha_3 P_{s-3}(z) \\ &= P_{s-2}(z) \left[\left(z - \frac{1}{2} \right) \left(z - \frac{1}{2} + \alpha_1 \right) + \alpha_2 - \tau_{s-1} \right] \\ &\quad - P_{s-3}(z) \left[\tau_{s-2} \left(z - \frac{1}{2} + \alpha_1 \right) - \alpha_3 \right]. \end{aligned}$$

Application of Lemma I.13.3 to the polynomials $f(z) = (z - \frac{1}{2})(z - \frac{1}{2} + \alpha_1) + \alpha_2 - \tau_{s-1}$ and $g(z) = \tau_{s-2}(z - \frac{1}{2} + \alpha_1) - \alpha_3$ shows that the corresponding quadrature formula has positive weights iff

$$\frac{\alpha_3}{\tau_{s-2}} \left(\alpha_1 - \frac{\alpha_3}{\tau_{s-2}} \right) - \alpha_2 + \tau_{s-1} > 0, \quad (13.14)$$

a result first found by Burrage (1978).

Necessary Conditions for Algebraic Stability

We now turn our attention to algebraic stability. We again use the notation $B(p)$, $C(\eta)$, $D(\xi)$ of Sect. IV.5.

Lemma 13.9 (Burrage 1982). *Consider Runge-Kutta methods, which satisfy $B(2)$ and the second condition for algebraic stability (i.e. M non-negative). Then,*

- a) $C(k)$ implies $B(2k-1)$;
- b) $D(k)$ implies $B(2k-1)$.

Proof. Instead of considering M , we work with the transformed matrix $\widehat{M} = V^T M V$ where $V = (c_i^{j-1})_{i,j=1}^s$ is the Vandermonde matrix. The elements of \widehat{M} are given by

$$\widehat{m}_{qr} = \sum_{i=1}^s b_i c_i^{q-1} \sum_{j=1}^s a_{ij} c_j^{r-1} + \sum_{j=1}^s b_j c_j^{r-1} \sum_{i=1}^s a_{ji} c_i^{q-1} - \sum_{i=1}^s b_i c_i^{q-1} \sum_{j=1}^s b_j c_j^{r-1}. \quad (13.15)$$

We further introduce

$$g_r = r \sum_{j=1}^s b_j c_j^{r-1} - 1$$

so that $B(\nu)$ is equivalent to $g_r = 0$ ($r = 1, \dots, \nu$). Then $C(k)$ simplifies (13.15) to

$$\hat{m}_{qr} = \frac{1}{q \cdot r} (g_{q+r} + 1 - (g_q + 1)(g_r + 1)) \quad q \leq k, r \leq k.$$

Similarly, $D(k)$ implies

$$\hat{m}_{qr} = -\frac{1}{q \cdot r} (g_{q+r} + g_q \cdot g_r) \quad q \leq k, r \leq k.$$

We now start with the hypothesis $B(2)$, i.e., $B(2l)$ for $l = 1$. This means that $g_1 = \dots = g_{2l} = 0$, so that, in both cases, $\hat{m}_{ll} = 0$. But if for a non-negative definite matrix a diagonal element is zero, the whole corresponding column must also be zero (see Exercise 11 of Sect. IV.12). This leads to $g_{l+q} = 0$ for $q = 1, \dots, k$; so we have $B(k+l)$. We then repeat the argument inductively until we arrive at $B(2k-1)$. \square

Since s -stage collocation methods satisfy $B(s)$ and $C(s)$ (see Theorem 7.8 of Chapter II) we have

Corollary 13.10 (Burrage 1978). *An s -stage algebraically stable collocation method must be of order at least $2s - 1$.* \square

Because *symmetric* methods have even order this gives:

Corollary 13.11 (Ascher & Bader 1986). *A symmetric algebraically stable collocation scheme has to be at Gaussian points.* \square

The next result states the necessity of the simplifying assumption $C(k)$. Observe that by Theorem 12.16 the weights b_i of DJ -irreducible, algebraically stable methods have to be positive.

Lemma 13.12. *If a Runge-Kutta method of order $p \geq 2k + 1$ satisfies $b_i > 0$ for $i = 1, \dots, s$, then the condition $C(k)$ holds.*

Proof (Dahlquist & Jeltsch (1979) attribute this idea to Butcher). The order conditions (see Sect. II.2)

$$\sum_{i=1}^s b_i c_i^{2q} = \frac{1}{2q+1}$$

$$\sum_{i,j=1}^s b_i c_i^q a_{ij} c_j^{q-1} = \frac{1}{(2q+1)q}$$

$$\sum_{i,j,m=1}^s b_i a_{ij} c_j^{q-1} a_{im} c_m^{q-1} = \frac{1}{(2q+1)q^2}$$

imply that

$$\sum_{i=1}^s b_i \left(\sum_{j=1}^s a_{ij} c_j^{q-1} - \frac{c_i^q}{q} \right)^2 = 0$$

for $2q+1 \leq p$. Since the b_i are positive, the individual terms of this sum must be zero for $q \leq k$. \square

A simple consequence of this lemma are the following *order barriers* for diagonally implicit DIRK ($a_{ij} = 0$ for $i < j$) and singly diagonally implicit SDIRK ($a_{ij} = 0$ for $i < j$ and $a_{ii} = \gamma$ for all i) methods.

Theorem 13.13 (Hairer 1980).

- a) A DIRK method with all b_i positive has order at most 6;
- b) An SDIRK method with all b_i positive has order at most 4;
- c) An algebraically stable DIRK method has order at most 4.

Proof. a) Suppose the order is greater than 6 and let i be the smallest index such that $c_i \neq 0$. Then by Lemma 13.12

$$a_{ii} c_i = \frac{c_i^2}{2}, \quad a_{ii} c_i^2 = \frac{c_i^3}{3},$$

contradicting $c_i \neq 0$.

- b) As above, we arrive for order greater than 4 at

$$a_{ii} c_i = \frac{c_i^2}{2} \quad \text{or} \quad a_{ii} = \frac{c_i}{2} (\neq 0).$$

Since for SDIRK methods we have $a_{ii} = a_{11}$, this leads to $c_1 = a_{11} \neq 0$, hence $i = 1$. Now $a_{11} = c_1/2$ contradicts $a_{11} = c_1$.

c) It is sufficient to consider DJ -irreducible methods, since the reduction process (see Table 12.3) leaves the class of DIRK methods invariant. From Theorem 12.16 and Lemma 13.12 we obtain that algebraic stability and order greater than 4 imply

$$a_{11} = c_1, \quad a_{11} c_1 = \frac{c_1^2}{2},$$

and hence $a_{11} = 0$. Inserted into m_{11} this yields $m_{11} = -b_1^2 < 0$, contradicting the non-negativity of the matrix M . \square

Similarly to Lemma 13.12 we have the following result for the second type of simplifying assumptions.

Lemma 13.14. *If a Runge-Kutta method of order $p \geq 2k + 1$ is algebraically stable and satisfies $b_i > 0$ for all i , then the condition $D(k)$ holds.*

Proof. The main idea is to use the W -transformation of Sect. IV.5 and to consider $W^T M W$ instead of M (see also the proof of Theorem 12.8). By Theorem 5.14 there exists a matrix W satisfying $T(k, k)$ (see Definition 5.10). With the help of Lemma 13.12 and Theorem 5.11a we obtain that the first k diagonal elements of

$$W^T M W = (W^T B W) X + X^T (W^T B W)^T - e_1 e_1^T \quad (13.16)$$

are zero. Since M and hence also $W^T M W$ is non-negative definite, the first k columns and rows of $W^T M W$ have to vanish. Thus the matrix $(W^T B W) X$ must be skew-symmetric in these regions (with exception of the first element). Because of $C(k)$ the first k columns and rows of $(W^T B W) X$ and X are identical. Thus the result follows from Theorem 5.11. \square

Characterization of Algebraically Stable Methods

Theorem 12.16, Lemma 13.12 and Lemma 13.14 imply that DJ -irreducible and algebraically stable RK-methods of order $p \geq 2k + 1$ satisfy $b_i > 0$ for all i , and the simplifying assumptions $C(k)$ and $D(k)$. These properties allow the following constructive characterization of all irreducible B -stable RK-methods.

Theorem 13.15 (Hairer & Wanner 1981). *Consider a p th order quadrature formula $(b_i, c_i)_{i=1}^s$ with positive weights and let W satisfy Property $T(k, k)$ of Definition 5.10 with $k = [(p - 1)/2]$. Then all p th order algebraically stable Runge-Kutta methods corresponding to this quadrature formula are given by*

$$A = W X W^{-1} \quad (13.17)$$

where

$$(W^T B W) X = \frac{1}{2} e_1 e_1^T + \begin{pmatrix} 0 & -\xi_1 & & \\ \xi_1 & \ddots & \ddots & \\ & \ddots & 0 & -\xi_k \\ & & \xi_k & \boxed{Q} \end{pmatrix} \quad (13.18)$$

and Q is an arbitrary matrix of dimension $s - k$ for which $Q + Q^T$ is non-negative definite. For p even we have to require that $q_{11} = 0$.

Proof. Algebraic stability and the positivity of the weights b_i imply $C(k)$ and $D(k)$ with $k = [(p - 1)/2]$. The matrix A of such a method can be written as

(13.17) with X given by (13.18). This follows from Theorem 5.11 and the fact that multiplication with $W^T B W$ does not change the first k columns and rows of X . This method is algebraically stable iff M (or $W^T M W$) is non-negative definite. By (13.16) this means that $Q + Q^T$ is non-negative definite.

Conversely, any Runge-Kutta method given by (13.17), (13.18) with $Q + Q^T$ non-negative definite is algebraically stable and satisfies $C(k)$ and $D(k)$. Therefore it follows from Theorem 5.1 in the case of odd $p = 2k + 1$ that the Runge-Kutta method is of order p .

If p is even, say $p = 2k + 2$, the situation is slightly more complicated. Because of

$$q_{11} = \sum_{i,j=1}^s b_i P_k(c_i) a_{ij} P_k(c_j)$$

it follows from $B(2k+2)$, $C(k)$, $D(k)$ that the order condition (13.19) below (with $\xi = \eta = k$) is equivalent to $q_{11} = 0$. The stated order p of the RK-method now follows from Lemma 13.16. \square

In the above proof we used the following modification of Theorem 5.1.

Lemma 13.16. *If the coefficients b_i, c_i, a_{ij} of an RK-method satisfy*

$$\sum_{i,j=1}^s b_i c_i^\xi a_{ij} c_j^\eta = \frac{1}{(\eta + \xi + 2)(\eta + 1)} \quad (13.19)$$

and $B(p)$, $C(\eta)$, $D(\xi)$ with $p \leq \eta + \xi + 2$ and $p \leq 2\eta + 2$, then the method is of order p .

Proof. The reduction process with the help of $C(\eta)$ and $D(\xi)$ as described in Sect. II.7 (Volume I) reduces all trees to the bushy trees covered by $B(p)$. The only exception is the tree corresponding to order condition (13.19). \square

Example 13.17 (Three-stage B -stable SIRK methods). Choose a third order quadrature formula with positive weights and let W satisfy $W^T B W = I$. Then (13.18) becomes

$$X = \begin{pmatrix} \frac{1}{2} & -\xi_1 & 0 \\ \xi_1 & a & b \\ 0 & c & d \end{pmatrix}, \quad \xi_1 = \frac{1}{2\sqrt{3}}.$$

The method is B -stable if $X^T + X - e_1 e_1^T$ is non-negative, i.e. if

$$a \geq 0, \quad d \geq 0, \quad 4ad \geq (c + b)^2. \quad (13.20)$$

If we want this method to be singly-implicit, we must have for the characteristic polynomial of A

$$\chi_A(z) = (1 - \gamma z)^3 = 1 - 3\gamma z + 3\gamma^2 z^2 - \gamma^3 z^3.$$

This means that (see (13.17))

$$\begin{aligned}\frac{1}{2} + a + d &= 3\gamma \\ \frac{a}{2} + \frac{1}{12} + \frac{d}{2} + ad - cb &= 3\gamma^2 \\ \frac{ad - cb}{2} + \frac{1}{12}d &= \gamma^3.\end{aligned}$$

Some elementary algebra shows that these equations can be solved and the inequalities (13.20) satisfied if $1/3 \leq \gamma \leq 1.06857902$, i.e., *exactly if* the corresponding rational approximation is A -stable (cf. Table 6.3; see also Hairer & Wanner (1981), where the analogous case with $s = p = 5$ is treated).

The “Equivalence” of A - and B -Stability

Many A -stable RK-methods are not B -stable (e.g., the trapezoidal rule, the Lobatto IIIA and Lobatto IIIB methods; see Theorem 12.12). On the other hand there is the famous result of Dahlquist (1978), saying that *every* A -stable *one-leg-method* is B -stable, which we shall prove in Sect. V.6. We have further seen in Example 13.17 that for a certain class of A -stable methods there is always a B -stable method with the same stability function. The general truth of this result was conjectured for many years and is as follows:

Theorem 13.18 (Hairer & Türke 1984, Hairer 1986). *Let $R(z) = P(z)/Q(z)$ ($P(0) = Q(0) = 1$, $\deg P \leq s$, $\deg Q = s$) be an irreducible, A -stable function satisfying $R(z) - e^z = \mathcal{O}(z^{p+1})$ for some $p \geq 1$. Then there exists an s -stage B -stable Runge-Kutta method of order p with $R(z)$ as stability function.*

Proof. Since $R(z)$ is an approximation to e^z of order p , it can be written in the form

$$R(z) = \frac{1 + \frac{1}{2}\Psi(z)}{1 - \frac{1}{2}\Psi(z)}, \quad \Psi(z) = \frac{z}{1} + \frac{\xi_1^2 z^2}{1} + \dots + \frac{\xi_{k-1}^2 z^2}{1} + \xi_k^2 z \Psi_k(z) \quad (13.21)$$

where $k = [(p-1)/2]$, $\xi_j^2 = 1/(4(4j^2 - 1))$ and $\Psi_k(z) = zg(z)/f(z)$ with $g(0) = f(0) = 1$, $\deg f \leq s - k$, $\deg g \leq s - k - 1$ (for p even we have in addition $g'(0) = f'(0)$). For the diagonal Padé-approximation $R^G(z)$ of order $2s$ this follows from Theorem 5.18 with $\nu = s - 1$ and $\Psi_\nu = z$:

$$R^G(z) = \frac{1 + \frac{1}{2}\Psi^G(z)}{1 - \frac{1}{2}\Psi^G(z)}, \quad \Psi^G(z) = \frac{z}{1} + \frac{\xi_1^2 z^2}{1} + \dots + \frac{\xi_{s-1}^2 z^2}{1}. \quad (13.22)$$

For an arbitrary $R(z)$ (satisfying the assumptions of the theorem) this is then a consequence of $R(z) = R^G(z) + \mathcal{O}(z^{p+1})$, or equivalently $\Psi(z) = \Psi^G(z) + \mathcal{O}(z^{p+1})$.

The function $R(z)$ of (13.21) is A -stable iff (Theorem 5.22)

$$\operatorname{Re} \Psi_k(z) < 0 \quad \text{for} \quad \operatorname{Re} z < 0.$$

Therefore, the function $\chi(z) = -\Psi_k(-1/z)$ is positive (c.f. Definition 5.19) and by Lemma 13.19 below there exists an $(s-k)$ -dimensional matrix Q such that

$$\chi(z) = e_1^T (Q + zI)^{-1} e_1 \quad \text{and} \quad Q + Q^T \quad \text{non-negative definite.}$$

We now fix an arbitrary quadrature formula of order p with positive weights b_i and (for the sake of simplicity) distinct nodes c_i . We let W be a matrix satisfying $W^T B W = I$ and Property $T(k, k)$ with $k = [(p-1)/2]$ (c.f. Lemma 5.12), and define the Runge-Kutta coefficients (a_{ij}) by (13.17) and (13.18). This Runge-Kutta method is algebraically stable, because $Q + Q^T$ is non-negative definite and of order p (observe that $g'(0) = f'(0)$ implies that the upper left element of Q vanishes). Finally, it follows from Theorem 5.18 and $\Psi_k(z) = -\chi(-1/z) = ze_1^T (I - zQ)^{-1} e_1$ that its stability function is $R(z)$. \square

It remains to prove the following lemma.

Lemma 13.19. *Let $\chi(z) = \alpha(z)/\beta(z)$ be an irreducible rational function with real polynomials*

$$\alpha(z) = z^{n-1} + \alpha_1 z^{n-2} + \dots, \quad \beta(z) = z^n + \beta_1 z^{n-1} + \dots \quad (13.23)$$

Then $\chi(z)$ is a positive function iff there exists an n -dimensional real matrix Q , such that

$$\chi(z) = e_1^T (Q + zI)^{-1} e_1 \quad \text{and} \quad Q + Q^T \quad \text{non-negative definite.} \quad (13.24)$$

Proof. a) The sufficiency follows from

$$\operatorname{Re} \chi(z) = q(z)^* \{ \operatorname{Re} z \cdot I + \tfrac{1}{2}(Q + Q^T) \} q(z)$$

with $q(z) = (Q + zI)^{-1} e_1$, since $Q + Q^T$ is non-negative definite.

b) For the proof of *necessity*, the hard part, we use Lemma 6.8 of Sect. V.6 below. This lemma is the essential ingredient for Dahlquist's equivalence result and will be proved in the chapter on multistep methods. It states that the positivity of $\chi(z)$ is equivalent to the existence of real, symmetric and non-negative definite matrices A and B , such that for arbitrary $z, w \in \mathbb{C}$ ($\vec{z} = (z^{n-1}, \dots, z, 1)^T$, $\vec{w} = (w^{n-1}, \dots, w, 1)^T$),

$$\alpha(z)\beta(w) + \alpha(w)\beta(z) = (z+w)\vec{z}^T A \vec{w} + \vec{z}^T B \vec{w}. \quad (13.25)$$

The matrix A is positive definite, if $\alpha(z)$ and $\beta(z)$ are relatively prime.

Comparing the coefficients of w^n in (13.25) we get

$$\alpha(z) = \vec{z}^T A e_1 \quad (13.26)$$

and observe that the first column of A consists of the coefficients of $\alpha(z)$. For the Cholesky decomposition of A , $A = U^T U$ (U is an upper triangular matrix) we

thus have $Ue_1 = e_1$. We next consider the possible computation of the matrix Q from the relation

$$(Q + zI)U\vec{z} = \beta(z) \cdot e_1 \quad (13.27)$$

or equivalently

$$QU\vec{z} = \beta(z) \cdot e_1 - zU\vec{z}. \quad (13.28)$$

The right-hand side of (13.28) is a known polynomial of degree $n-1$, since $Ue_1 = e_1$. Therefore, a comparison of the coefficients in (13.28) yields the matrix QU and hence also Q . It remains to prove that this matrix Q satisfies (13.24).

Using (13.27), the formula $Ae_1 = U^T U e_1 = U^T e_1$ and (13.26) we obtain

$$e_1^T (Q + zI)^{-1} e_1 \cdot \beta(z) = e_1^T U\vec{z} = e_1^T A^T \vec{z} = \alpha(z), \quad (13.29)$$

which verifies the first relation of (13.24). Further, from (13.27) and $\alpha(z) = e_1^T U\vec{z}$ we get

$$\vec{z}^T U^T (Q + wI) U \vec{w} = \alpha(z) \beta(w).$$

Inserting this formula and the analogous one (with z and w exchanged) into (13.25) yields $0 = \vec{z}^T (B - U^T (Q + Q^T) U) \vec{w}$, so that $B = U^T (Q + Q^T) U$. This verifies the second relation of (13.24), since B is symmetric and non-negative definite. \square

Exercises

1. (Perron (1913) attributes this result to Wallis, *Arithmetica infinitorum* 1655 and Euler 1737). Let the sequences $\{A_k\}$ and $\{B_k\}$ be given by

$$\begin{aligned} A_k &= b_k A_{k-1} + a_k A_{k-2}, & A_{-1} &= 1, & A_0 &= b_0 \\ B_k &= b_k B_{k-1} + a_k B_{k-2}, & B_{-1} &= 0, & B_0 &= 1 \end{aligned} \quad (13.30)$$

then

$$\frac{A_n}{B_n} = b_0 + \frac{a_1}{b_1} + \dots + \frac{a_n}{b_n}. \quad (13.31)$$

Hint. Let $x = (x_0, x_1, \dots, x_{n+1})^T$ be the solution of $Mx = (0, \dots, 0, 1)^T$, where

$$M = \begin{pmatrix} 1 & -b_0 & -a_1 & & & \\ & 1 & -b_1 & -a_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -b_{n-1} & -a_n \\ & & & & 1 & -b_n \\ & & & & & 1 \end{pmatrix}.$$

One easily finds

$$\frac{x_0}{x_1} = b_0 + \frac{a_1}{x_1/x_2} = b_0 + \frac{a_1}{b_1} + \frac{a_2}{x_2/x_3} = \dots$$

so that x_0/x_1 is equal to the right hand side of (13.31). The statement now follows from the fact that

$$\begin{aligned}(A_{-1}, A_0, \dots, A_n)M &= (1, 0, \dots, 0) \\ (B_{-1}, B_0, \dots, B_n)M &= (0, 1, 0, \dots, 0).\end{aligned}$$

implying $x_0 = A_n$ and $x_1 = B_n$.

2. Let $P_s(z)$ be the Legendre polynomial (13.4) and $N_s^G(z)$ defined by the recursion (13.5) with $N_0^G(z) = 0$, $N_1^G(z) = 1$. Prove that

$$N_{s-k}^G(z)P_{s-k-1}(z) - N_{s-k-1}^G(z)P_{s-k}(z) = \tau_1 \cdot \tau_2 \cdot \dots \cdot \tau_{s-k-1}.$$

Hint. Use the relation

$$\begin{pmatrix} N_m^G(z) & P_m(z) \\ N_{m-1}^G(z) & P_{m-1}(z) \end{pmatrix} = \begin{pmatrix} z - \frac{1}{2} & -\tau_{m-1} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} N_{m-1}^G(z) & P_{m-1}(z) \\ N_{m-2}^G(z) & P_{m-2}(z) \end{pmatrix}.$$

3. Consider the Hermitian quadrature formula

$$\int_0^1 f(x)dx = b_1 f(c_1) + \alpha f(c_2) + \beta \frac{f'(c_2)}{1!} + \gamma \frac{f''(c_2)}{2!}. \quad (13.32)$$

Replace $f'(c_2)$ and $f''(c_2)$ by finite divided differences based on $f(c_2 - \varepsilon)$, $f(c_2)$, $f(c_2 + \varepsilon)$ to obtain a quadrature formula

$$\int_0^1 f(x)dx = \bar{b}_1 f(c_1) + \bar{b}_2 f(c_2 - \varepsilon) + \bar{b}_3 f(c_2) + \bar{b}_4 f(c_2 + \varepsilon). \quad (13.33)$$

a) Compute $Q(z)$ for Formula (13.33) and obtain, by letting $\varepsilon \rightarrow 0$, an expression which generalizes (13.2) to Hermitian quadrature formulas.

b) Compute the values of b_1 and b_2 ($l_1 = 1, l_2 = 3$) of (13.10').

c) Show that $n_+ - n_-$ (see Theorem 13.5) is the same for (13.32) and (13.33) with ε sufficiently small.

Results. a)
$$Q(z) = \frac{b_1}{z - c_1} + \frac{\alpha}{z - c_2} + \frac{\beta}{(z - c_2)^2} + \frac{\gamma}{(z - c_2)^3}$$

b)
$$b_1 = b_1 \quad (\text{sic!}), \quad b_2 = \gamma/3!.$$

4. The rational function $\chi(z) = \alpha(z)/\beta(z)$ with $\alpha(z) = z + \alpha_1$, $\beta(z) = z^2 + \beta_1 z + \beta_2$ is positive, iff $\alpha_1 \geq 0$, $\beta_2 \geq 0$, $\beta_1 - \alpha_1 \geq 0$ (compare (5.48))

a) Find real, symmetric and non-negative definite matrices A and B such that (13.25) holds.

b) Show that these matrices are, in general, not unique.

c) As in the proof of Lemma 13.19, compute the matrix Q such that (13.24) holds.

Hint. Begin with the construction of B by putting $w = -z$ in (13.25).

IV.14 Existence and Uniqueness of IRK Solutions

Jusqu'à présent, nous avons supposé que le schéma admettait une solution. Pour en démontrer l'existence ...

(Crouzeix & Raviart 1980)

Since contractivity without feasibility makes little sense ...

(M.N. Spijker 1985)

Since the Runge-Kutta methods studied in the foregoing sections are all implicit, we have to ensure that the numerical solutions, for which we have derived so many nice results, also really exist. The existence theory for implicit Runge-Kutta methods, presented in Volume I (Theorem II.7.2), is for the non-stiff case only, where hL is small (L the Lipschitz constant). This is not a reasonable assumption for the stiff case.

We shall study the existence of a Runge-Kutta solution, defined implicitly by

$$g_i = y_0 + h \sum_{j=1}^s a_{ij} f(x_0 + c_j h, g_j), \quad i = 1, \dots, s \quad (14.1a)$$

$$y_1 = y_0 + h \sum_{j=1}^s b_j f(x_0 + c_j h, g_j), \quad (14.1b)$$

for differential equations which satisfy the one-sided Lipschitz condition

$$\langle f(x, y) - f(x, z), y - z \rangle \leq \nu \|y - z\|^2. \quad (14.2)$$

Existence

It was first pointed out by Crouzeix & Raviart (1980) that the coercivity of the Runge-Kutta matrix A (or of its inverse) plays an important role for the proof of existence.

Definition 14.1. We consider the inner product $\langle u, v \rangle_D = u^T D v$, where $D = \text{diag}(d_1, \dots, d_s)$ with $d_i > 0$. We then denote by $\alpha_D(A^{-1})$ the largest number α such that

$$\langle u, A^{-1} u \rangle_D \geq \alpha \langle u, u \rangle_D \quad \text{for all } u \in \mathbb{R}^s. \quad (14.3)$$

We also set

$$\alpha_0(A^{-1}) = \sup_{D>0} \alpha_D(A^{-1}). \quad (14.4)$$

The first existence results for the above problem were given by Crouzeix & Raviart (1980), Dekker (1982) and Crouzeix, Hundsdorfer & Spijker (1983). Their results can be summarized as follows:

Theorem 14.2. *Let f be continuously differentiable and satisfy (14.2). If the Runge-Kutta matrix A is invertible and*

$$h\nu < \alpha_0(A^{-1}) \quad (14.5)$$

then the nonlinear system (14.1a) possesses a solution (g_1, \dots, g_s) .

Proof. The original proofs are based on the “uniform monotonicity theorem” or on similar results. We present here a more elementary version which, however, has the disadvantage of requiring the differentiability hypothesis for f . The idea is to consider the homotopy

$$g_i = y_0 + h \sum_{j=1}^s a_{ij} f(x_0 + c_j h, g_j) + (\tau - 1)h \sum_{j=1}^s a_{ij} f(x_0 + c_j h, y_0), \quad (14.6)$$

which is constructed in such a way that for $\tau = 0$ the system (14.6) has the solution $g_i = y_0$, and for $\tau = 1$ it is equivalent to (14.1a). We consider g_i as functions of τ and differentiate (14.6) with respect to this parameter. This gives

$$\dot{g}_i = h \sum_{j=1}^s a_{ij} \frac{\partial f}{\partial y}(x_0 + c_j h, g_j) \cdot \dot{g}_j + h \sum_{j=1}^s a_{ij} f(x_0 + c_j h, y_0)$$

or equivalently

$$(I - h(A \otimes I)\{f_y\}) \dot{g} = h(A \otimes I)f_0 \quad (14.7)$$

where we have used the notations

$$\dot{g} = (\dot{g}_1, \dots, \dot{g}_s)^T, \quad f_0 = (f(x_0 + c_1 h, y_0), \dots, f(x_0 + c_s h, y_0))^T$$

(more precisely, \dot{g} should be written as $(\dot{g}_1^T, \dots, \dot{g}_s^T)^T$) and

$$\{f_y\} = \text{blockdiag} \left(\frac{\partial f}{\partial y}(x_0 + c_1 h, g_1), \dots, \frac{\partial f}{\partial y}(x_0 + c_s h, g_s) \right).$$

In order to show that \dot{g} can be expressed as $\dot{g} = G(g)$ with a globally bounded $G(g)$, we take a D satisfying $h\nu < \alpha_D(A^{-1})$, multiply (14.7) by $\dot{g}^T(DA^{-1} \otimes I)$ and so obtain

$$\dot{g}^T(DA^{-1} \otimes I)\dot{g} - h\dot{g}^T(D \otimes I)\{f_y\}\dot{g} = h\dot{g}^T(D \otimes I)f_0. \quad (14.8)$$

We now estimate the three individual terms of this equation.

The estimate

$$\dot{g}^T(DA^{-1} \otimes I)\dot{g} \geq \alpha_D(A^{-1}) \|\dot{g}\|_D^2, \quad (14.9)$$

where we have introduced the notation $\|\dot{g}\|_D^2 = \dot{g}^T(D \otimes I)\dot{g}$, is (14.3) in the case of *scalar* differential equations (absence of “ $\otimes I$ ”). In the general case we must apply the ideas of Exercise 1 of Sect. IV.12 to the matrix $\frac{1}{2}(DA^{-1} + (DA^{-1})^T) - \alpha_D(A^{-1})D$, which is non-negative definite by Definition 14.1. It follows from (14.2) with $y = z + \varepsilon u$ that

$$\left\langle \varepsilon \frac{\partial f}{\partial y}(x, z)u + o(\varepsilon), \varepsilon u \right\rangle \leq \nu \varepsilon^2 \|u\|^2.$$

Dividing by ε^2 and taking the limit $\varepsilon \rightarrow 0$ we obtain $\langle u, \frac{\partial f}{\partial y}(x, z)u \rangle \leq \nu \|u\|^2$ for all (x, z) and all u . Consequently we also have

$$\dot{g}^T(D \otimes I)\{f_y\}\dot{g} \leq \nu \|\dot{g}\|_D^2. \quad (14.10)$$

The right-hand term of (14.8) is bounded by $h \|\dot{g}\|_D \cdot \|f_0\|_D$ by the Cauchy-Schwarz-Bunjakowski inequality.

Inserting these three estimates into (14.8) yields

$$(\alpha_D(A^{-1}) - h\nu) \|\dot{g}\|_D^2 \leq h \|\dot{g}\|_D \cdot \|f_0\|_D.$$

This proves that \dot{g} can be written as $\dot{g} = G(g)$ with

$$\|G(g)\|_D \leq \frac{h \|f_0\|_D}{\alpha_D(A^{-1}) - h\nu}.$$

It now follows from Theorem 7.4 (Sect. I.7) that this differential equation with initial values $g_i(0) = y_0$ possesses a solution for all τ , in particular also for $\tau = 1$. This proves the existence of a solution of (14.1a). \square

Remark. It has recently been shown by Kraaijevanger & Schneid (1991, Theorem 2.12) that Condition (14.5) is “essentially optimal”.

A Counterexample

After our discussion that Monday afternoon (October 1980) I went for a walk and I got the idea for the counterexample.
(M.N. Spijker)

The inequality in (14.5) is *strict*, therefore Theorem 14.2 (together with Exercise 1 below) does not yet answer the simple question: “does a B -stable method on a contractive problem ($\nu = 0$) always admit a solution”. A first counterexample to this statement has been given by Crouzeix, Hundsdorfer & Spijker (1983). An easy idea for constructing another counterexample is to use the W -transformation (see Sections IV.5 and IV.13) as follows:

We put $s = 4$ and take a quadrature formula with positive weights, say,

$$(c_i) = (0, 1/3, 2/3, 1), \quad (b_i) = (1/8, 3/8, 3/8, 1/8).$$

We then construct a matrix W satisfying property $T(1, 1)$ according to Lemma 5.12. This yields for the above quadrature formula

$$W = \begin{pmatrix} 1 & -\sqrt{3} & \sqrt{3} & -1 \\ 1 & -\sqrt{3}/3 & -\sqrt{3}/3 & 1 \\ 1 & \sqrt{3}/3 & -\sqrt{3}/3 & -1 \\ 1 & \sqrt{3} & \sqrt{3} & 1 \end{pmatrix}.$$

Finally, we put (with $\xi_1 = 1/(2\sqrt{3})$)

$$A = W X W^{-1} \quad \text{with} \quad X = \begin{pmatrix} 1/2 & -\xi_1 & 0 & 0 \\ \xi_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\beta \\ 0 & 0 & \beta & 0 \end{pmatrix}.$$

For $\beta = 1/(4\sqrt{3})$ this gives nice rational coefficients for the RK-matrix, namely

$$A = \frac{1}{48} \begin{pmatrix} 3 & 0 & 3 & -6 \\ 6 & 9 & 0 & 1 \\ 5 & 18 & 9 & 0 \\ 12 & 15 & 18 & 3 \end{pmatrix}.$$

It follows from Theorem 13.15 that this method is algebraically stable and of order 4. However, $\pm i\beta$ is an eigenvalue pair of X and hence also of A .

We thus choose the differential equation

$$y' = Jy + f(x) \quad \text{with} \quad J = \begin{pmatrix} 0 & -1/\beta \\ 1/\beta & 0 \end{pmatrix},$$

which satisfies (14.2) with $\nu = 0$ independent of the choice of $f(x)$. If we apply the above method with $h = 1$ to this problem and initial values $x_0 = 0$, $y_0 = (0, 0)^T$, Eq. (14.1a) becomes equivalent to the linear system

$$(I - A \otimes J)g = (A \otimes I)f_0,$$

where $g = (g_1, \dots, g_4)^T$ and $f_0 = (f(c_1), \dots, f(c_4))^T$. The matrix $(I - A \otimes J)$ is singular because the eigenvalues of $I - A \otimes J$ are just $1 - \lambda\mu$ where λ and μ are the eigenvalues of A and J , respectively. However, A is regular, therefore it is possible to choose $f(x)$ in such a way that this equation does not have a solution.

Influence of Perturbations and Uniqueness

Our next problem is the question, how *perturbations* in the Runge-Kutta equations influence the numerical solution. Research into this problem was initiated independently by Frank, Schneid & Ueberhuber (preprint 1981, published 1985) and Dekker (1982).

As above, we use the notations

$$\begin{aligned} \|u\|_D &= \sqrt{u^T D u} = \sqrt{\langle u, u \rangle_D} & u \in \mathbb{R}^s \\ \|g\|_D &= \sqrt{g^T (D \otimes I) g} & g \in \mathbb{R}^{sn} \end{aligned}$$

and $\|A\|_D$ for the corresponding matrix norm.

Theorem 14.3 (Dekker 1984). *Let g_i and y_1 be given by (14.1) and consider perturbed values \hat{g}_i and \hat{y}_1 satisfying*

$$\hat{g}_i = y_0 + h \sum_{j=1}^s a_{ij} f(x_0 + c_j h, \hat{g}_j) + \delta_i \quad (14.11a)$$

$$\hat{y}_1 = y_0 + h \sum_{j=1}^s b_j f(x_0 + c_j h, \hat{g}_j). \quad (14.11b)$$

If the Runge-Kutta matrix A is invertible, if the one-sided Lipschitz condition (14.2) is satisfied, and $h\nu < \alpha_D(A^{-1})$ for some positive diagonal matrix D , then we have the estimates

$$\|\hat{g} - g\|_D \leq \frac{\|A^{-1}\|_D}{\alpha_D(A^{-1}) - h\nu} \|\delta\|_D \quad (14.12)$$

$$\|\hat{y}_1 - y_1\| \leq \|b^T A^{-1}\|_D \left(1 + \frac{\|A^{-1}\|_D}{\alpha_D(A^{-1}) - h\nu} \right) \|\delta\|_D, \quad (14.13)$$

where $g = (g_1, \dots, g_s)^T$, $\hat{g} = (\hat{g}_1, \dots, \hat{g}_s)^T$, and $\delta = (\delta_1, \dots, \delta_s)^T$.

Proof. With the notation $\Delta g = \hat{g} - g$ and

$$\Delta f = \left(f(x_0 + c_1 h, \hat{g}_1) - f(x_0 + c_1 h, g_1), \dots, f(x_0 + c_s h, \hat{g}_s) - f(x_0 + c_s h, g_s) \right)^T$$

the difference of (14.11a) and (14.1a) can be written as

$$\Delta g = h(A \otimes I) \Delta f + \delta.$$

As in the proof of Theorem 14.2 we multiply this equation by $\Delta g^T(DA^{-1} \otimes I)$ and obtain

$$\Delta g^T(DA^{-1} \otimes I) \Delta g - h \Delta g^T(D \otimes I) \Delta f = \Delta g^T(DA^{-1} \otimes I) \delta. \quad (14.14)$$

This equation is very similar to Eq. (14.8) and we estimate it in the same way: since D is a diagonal matrix with positive entries, it follows from (14.2) that

$$\Delta g^T(D \otimes I) \Delta f \leq \nu \|\Delta g\|_D^2. \quad (14.15)$$

Inserting (14.15) and (14.9) (with \dot{g} replaced by Δg) into (14.14) we get

$$(\alpha_D(A^{-1}) - h\nu) \|\Delta g\|_D^2 \leq \|\Delta g\|_D \|(A^{-1} \otimes I) \delta\|_D$$

which implies (14.12). The estimate (14.13) then follows immediately from

$$\hat{y}_1 - y_1 = h(b^T \otimes I) \Delta f = (b^T A^{-1} \otimes I)(\Delta g - \delta). \quad \square$$

Putting $\delta = 0$ in Theorem 14.3 we get the following uniqueness result.

Theorem 14.4. *Consider a differential equation satisfying (14.2). If the Runge-Kutta matrix A is invertible and $h\nu < \alpha_0(A^{-1})$, then the system (14.1a) possesses at most one solution.* \square

Computation of $\alpha_0(A^{-1})$

... the determination of a suitable matrix D ... This task does not seem easy at first glance ... (K. Dekker 1984)

The value $\alpha_D(A^{-1})$ of Definition 14.1 is the smallest eigenvalue of the symmetric matrix $(D^{1/2}A^{-1}D^{-1/2} + (D^{1/2}A^{-1}D^{-1/2})^T)/2$. The computation of $\alpha_0(A^{-1})$ is more difficult, because the optimal D is not known in general.

An upper bound for $\alpha_0(A^{-1})$ is

$$\alpha_0(A^{-1}) \leq \min_{i=1,\dots,s} \omega_{ii} \quad (14.16)$$

where ω_{ij} are the entries of A^{-1} . This follows from (14.3) by putting $u = e_i$, the i th unit vector.

Lower bounds for $\alpha_0(A^{-1})$ were first given by Frank, Schneid & Ueberhuber in 1981. Following are the exact values due to Dekker (1984), Dekker & Verwer (1984, p. 55-164), and Dekker & Hairer (1985) (see also Liu & Kraaijevanger 1988 and Kraaijevanger & Schneid 1991).

Theorem 14.5. *For the methods of Sect. IV.5 we have:*

Gauss	$\alpha_0(A^{-1}) = \min_{i=1,\dots,s} \frac{1}{2c_i(1-c_i)},$
Radau IA	$\alpha_0(A^{-1}) = \begin{cases} 1 & \text{if } s = 1, \\ \frac{1}{2(1-c_2)} & \text{if } s > 1, \end{cases}$
Radau IIA	$\alpha_0(A^{-1}) = \begin{cases} 1 & \text{if } s = 1, \\ \frac{1}{2c_{s-1}} & \text{if } s > 1, \end{cases}$
Lobatto IIIC	$\alpha_0(A^{-1}) = \begin{cases} 1 & \text{if } s = 2, \\ 0 & \text{if } s > 2. \end{cases}$

Proof. a) Gauss methods: written out in “symmetricized form”, estimate (14.3) reads

$$\frac{1}{2} u^T (DA^{-1} + (DA^{-1})^T) u \geq \alpha u^T D u.$$

Evidently the sharpest estimates come out if D is such that the left-hand matrix is as “close to diagonal as possible”. After many numerical computations, Dekker had the nice surprise that with the choice $D = B(C^{-1} - I)$, where $B = \text{diag}(b_1, \dots, b_s)$ and $C = \text{diag}(c_1, \dots, c_s)$, the matrix

$$DA^{-1} + (DA^{-1})^T = BC^{-2} \quad (14.17)$$

becomes completely diagonal. Then the optimal α is simply obtained by testing the unit vectors $u = e_k$, which gives

$$\alpha_0(A^{-1}) = \min_i \frac{b_i}{2c_i^2 d_i} = \min_i \frac{b_i}{2c_i^2 b_i(1/c_i - 1)} = \min_i \frac{1}{2c_i(1 - c_i)}.$$

It remains to prove (14.17): we verify the equivalent formula

$$V^T(A^T D + DA - A^T BC^{-2}A)V = 0 \quad (14.18)$$

where $V = (c_i^{j-1})$ is the Vandermonde matrix. The (l, m) -element of the matrix (14.18) is

$$\begin{aligned} \sum_{i,j} b_j \left(\frac{1}{c_j} - 1 \right) a_{ji} c_i^{l-1} c_j^{m-1} + \sum_{i,j} b_i \left(\frac{1}{c_i} - 1 \right) a_{ij} c_i^{l-1} c_j^{m-1} \\ - \sum_{i,j,k} b_i \frac{1}{c_i^2} a_{ik} c_k^{l-1} a_{ij} c_j^{m-1}. \end{aligned} \quad (14.19)$$

With the help of the simplifying assumptions $C(s)$ and $B(2s)$ the expression (14.19) can be seen to be zero.

b) For the Radau IA methods we take $D = B(I - C)$ and show that

$$DA^{-1} + (DA^{-1})^T = B + e_1 e_1^T. \quad (14.20)$$

The stated formula for $\alpha_0(A^{-1})$ then follows from $0 = c_1 < c_2 < \dots < c_s$ and from

$$\frac{b_1 + 1}{b_1} \geq \frac{1}{1 - c_2},$$

which is a simple consequence of $b_1 = 1/s^2$ (see Abramowitz & Stegun (1964), Formula 25.4.31). For the verification of (14.20) one shows that $V^T(DA^{-1} + (DA^{-1})^T - B - e_1 e_1^T)V = 0$. Helpful formulas for this verification are $A^{-1}V e_1 = b_1^{-1} e_1$, $V^T e_1 = e_1$ and $A^{-1}V e_j = (j-1)V e_{j-1}$ for $j \geq 2$.

c) Similarly, the statement for the Radau IIA methods follows with $D = BC^{-1}$ from the identity

$$DA^{-1} + (DA^{-1})^T = BC^{-2} + e_s e_s^T.$$

d) As in part (b) one proves for the Lobatto IIIC methods that

$$BA^{-1} + (BA^{-1})^T = e_1 e_1^T + e_s e_s^T. \quad (14.21)$$

Since this matrix is diagonal, we obtain $\alpha_0(A^{-1}) = 1$ for $s = 2$ and $\alpha_0(A^{-1}) = 0$ for $s > 2$. \square

For diagonally implicit Runge-Kutta methods we have the following result.

Theorem 14.6 (Montijano 1983). *For a DIRK-method with positive a_{ii} we have*

$$\alpha_0(A^{-1}) = \min_{i=1, \dots, s} \frac{1}{a_{ii}}. \quad (14.22)$$

Proof. With $D = \text{diag}(1, \varepsilon^2, \varepsilon^4, \dots, \varepsilon^{2s-2})$ we obtain

$$D^{1/2} A^{-1} D^{-1/2} + (D^{1/2} A^{-1} D^{-1/2})^T = \text{diag}(a_{11}^{-1}, \dots, a_{ss}^{-1}) + \mathcal{O}(\varepsilon),$$

so that $\alpha_0(A^{-1}) \geq \min_i a_{ii}^{-1} + \mathcal{O}(\varepsilon)$. This inequality for $\varepsilon \rightarrow 0$ and (14.16) prove the statement. \square

Methods with Singular A

For the Lobatto IIIA methods the first stage is explicit (the first row of A vanishes) and for the Lobatto IIIB methods the last stage is explicit (the last column of A vanishes). For these methods the Runge-Kutta matrix is of the form

$$A = \begin{pmatrix} 0 & 0 \\ a & \tilde{A} \end{pmatrix} \quad \text{or} \quad A = \begin{pmatrix} \tilde{A} & 0 \\ a^T & 0 \end{pmatrix} \quad (14.23)$$

and we have the following variant of Theorem 14.2.

Theorem 14.7. *Let f be continuously differentiable and satisfy (14.2). If the Runge-Kutta matrix is given by one of the matrices in (14.23) with invertible \tilde{A} , then the assumption*

$$h\nu < \alpha_0(\tilde{A}^{-1})$$

implies that the nonlinear system (14.1a) has a solution.

Proof. The explicit stage poses no problem for the existence of a solution. To obtain the result we repeat the proof of Theorem 14.2 for the $s - 1$ implicit stages (i.e., A is replaced by \tilde{A} and the inhomogeneity in (14.6) may be different). \square

An explicit formula for $\alpha_0(\tilde{A}^{-1})$ for the Lobatto IIIB methods has been given by Dekker & Verwer (1984), and for the Lobatto IIIA methods by Liu, Dekker & Spijker (1987). The result is

Theorem 14.8. *We have for*

$$\begin{array}{ll} \text{Lobatto IIIA} & \alpha_0(\tilde{A}^{-1}) = \begin{cases} 2 & \text{if } s = 2, \\ c_{s-1}^{-1} & \text{if } s > 2, \end{cases} \\ \text{Lobatto IIIB} & \alpha_0(\tilde{A}^{-1}) = \begin{cases} 2 & \text{if } s = 2, \\ (1 - c_2)^{-1} & \text{if } s > 2. \end{cases} \end{array}$$

Proof. For the Lobatto IIIA methods we put $D = \tilde{B}\tilde{C}^{-2}$ with the diagonal matrices $\tilde{B} = \text{diag}(b_2, \dots, b_s)$ and $\tilde{C} = \text{diag}(c_2, \dots, c_s)$. As in part (a) of the proof of Theorem 14.5 we get

$$D\tilde{A}^{-1} + (D\tilde{A}^{-1})^T = e_{s-1}e_{s-1}^T + 2\tilde{B}\tilde{C}^{-3}$$

which implies the formula for $\alpha_0(\tilde{A}^{-1})$, because $b_s = (s(s-1))^{-1}$ and $(1 + 2b_s) \geq b_s/c_{s-1}$ for $s > 2$.

For the Lobatto IIIB methods the choice $D = \tilde{B}(I - \tilde{C})^2$ (with the matrices $\tilde{B} = \text{diag}(b_1, \dots, b_{s-1})$, $\tilde{C} = \text{diag}(c_1, \dots, c_{s-1})$) leads to

$$D\tilde{A}^{-1} + (D\tilde{A}^{-1})^T = e_1e_1^T + 2\tilde{B}(I - \tilde{C}).$$

This proves the second statement. \square

Methods with explicit stages (such as Lobatto IIIA and IIIB) don't allow estimates of the numerical solution in the presence of arbitrary perturbations. They are usually not AN -stable and $K(Z)$ is not bounded (see Theorem 12.12). Nevertheless we have the following uniqueness result.

Theorem 14.9. *Consider a differential equation satisfying (14.2). If the Runge-Kutta matrix is of the form (14.23) with invertible \tilde{A} and if $h\nu < \alpha_0(\tilde{A}^{-1})$, then the nonlinear system (14.1a) has at most one solution.*

Proof. Suppose, there exists a second solution \hat{g}_i satisfying (14.11a) with $\delta_i = 0$.

a) If the first stage is explicit we have $\hat{g}_1 = g_1$. The difference of the two Runge-Kutta formulas then yields

$$\Delta g = h(\tilde{A} \otimes I)\Delta f$$

with $\Delta g = (\hat{g}_i - g_i)_{i=2}^s$ and $\Delta f = (f(x_0 + c_i h, \hat{g}_i) - f(x_0 + c_i h, g_i))_{i=2}^s$. As in the proof of Theorem 14.3 we then conclude that $\Delta g = 0$.

b) In the second case we can apply Theorem 14.3 to the first $s-1$ stages, which yields uniqueness of g_1, \dots, g_{s-1} . Clearly, g_s also is unique, because the last stage is explicit. \square

Lobatto IIIC Methods

For the Lobatto IIIC methods with $s \geq 3$ we have $\alpha_0(A^{-1}) = 0$ (see Theorem 14.5). Since these methods are algebraically stable it is natural to ask whether the nonlinear system (14.1a) also has a solution for differential equations satisfying (14.2) with $\nu = 0$. A positive answer to this question has been given by Hundsdorfer & Spijker (1987) for the case $s = 3$, and by Liu & Kraaijevanger (1988) for the general case $s \geq 3$ (see Exercise 6 below; see also Kraaijevanger & Schneid 1991).

Exercises

1. Prove that $\alpha_0(A) \geq 0$ for algebraically stable Runge-Kutta methods. Also, $\alpha_0(A^{-1}) \geq 0$ if in addition the matrix A is invertible.
2. Let A be a real matrix. Show that $\alpha_0(A) \leq \operatorname{Re} \lambda$, where λ is an eigenvalue of A .
3. (Hundsdorfer 1985, Cooper 1986). Prove that Theorem 14.2 remains valid for singular A , if $h\nu < \alpha$ with α satisfying

$$\langle u, Au \rangle_D \geq \alpha \langle Au, Au \rangle_D \quad \text{for all } u \in \mathbb{R}^s.$$

Hint. Use the transformation $g = \mathbb{1} \otimes y_0 + (A \otimes I)k$ and apply the ideas of the proof of Theorem 14.2 to the homotopy

$$k_i = f(x_0 + c_i h, y_0 + h \sum_{j=1}^s a_{ij} k_j) + (\tau - 1)f(x_0 + c_i h, y_0).$$

4. (Barker, Berman & Plemmons 1978, Montijano 1983). Prove that for any two-stage method the condition

$$a_{11} > 0, \quad a_{22} > 0, \quad \det(A) > 0 \quad (14.24)$$

is equivalent to $\alpha_0(A^{-1}) > 0$.

Remark. For a generalization of this result to three-stage methods see Kraaijevanger (1991).

5. For the two-stage Radau IIA method we have $\alpha_0(A^{-1}) = 3/2$. Construct a differential equation $y' = \lambda(x)y$ with $\operatorname{Re} \lambda(x) = 3/2 + \varepsilon$ ($\varepsilon > 0$ arbitrarily small) such that the Runge-Kutta equations do not admit a unique solution for all $h > 0$.
6. Prove that for the Lobatto IIIC methods (with $s \geq 3$) the matrix

$$I - (A \otimes I)J \quad \text{with} \quad J = \operatorname{blockdiag}(J_1, \dots, J_s)$$

is non-singular, if $\mu_2(J_k) \leq 0$. This implies that the Runge-Kutta equations (14.1a) have a unique solution for all problems $y' = A(x)y + f(x)$ with $\mu_2(A(x)) \leq 0$.

Hint (Liu & Kraaijevanger 1988, Liu, Dekker & Spijker 1987). Let $v = (v_1, \dots, v_s)^T$ be a solution of $(I - (A \otimes I)J)v = 0$. With the help of (14.21) show first that $v_1 = v_s = 0$. Then consider the $(s-2)$ -dimensional submatrix $\tilde{A} = (a_{ij})_{i,j=2}^{s-1}$ and prove $\alpha_0(\tilde{A}^{-1}) > 0$ by considering the diagonal matrix $\tilde{D} = \operatorname{diag}(b_i(c_i^{-1} - 1)^2)_{i=2}^{s-1}$.

7. Consider an algebraically stable Runge-Kutta method with invertible A and apply it to the differential equation $y' = (J(x) - \varepsilon I)y + f(x)$ where $\mu(J(x)) \leq 0$ and $\varepsilon > 0$. Prove that the numerical solution $y_1(\varepsilon)$ converges to a limit for $\varepsilon \rightarrow 0$, whereas the internal stages $g_i(\varepsilon)$ need not converge.

Hint. Expand the $g_i(\varepsilon)$ in a series $g_i(\varepsilon) = \varepsilon^{-1}g_i^{(-1)} + g_i^{(0)} + \varepsilon g_i^{(1)} + \dots$ and prove the implication

$$g = (A \otimes I)Jg \implies (b^T \otimes I)Jg = 0$$

where $J = \operatorname{blockdiag}(J(x_0 + c_1 h), \dots, J(x_0 + c_s h))$.

IV.15 B-Convergence

In using A -stable one-step methods to solve large systems of stiff nonlinear differential equations, we have found that

- (a) some A -stable methods give highly unstable solutions, and
- (b) the accuracy of the solutions obtained when the equations are stiff often appears to be unrelated to the order of the method used.

This has caused us to re-examine the form of stability required when stiff systems of equations are solved, and to question the relevance of the concept of (nonstiff) order of accuracy for stiff problems.
(A. Prothero & A. Robinson 1974)

Prothero & Robinson (1974) were the first to discover the order reduction of implicit Runge-Kutta methods when applied to stiff differential equations. Frank, Schneid & Ueberhuber (1981) then introduced the “concept of B -convergence”, which furnishes global error estimates independent of the stiffness.

The Order Reduction Phenomenon

For the study of the accuracy of Runge-Kutta methods applied to stiff differential equations, Prothero & Robinson (1974) proposed considering the problem

$$y' = \lambda(y - \varphi(x)) + \varphi'(x), \quad y(x_0) = \varphi(x_0), \quad \operatorname{Re} \lambda \leq 0. \quad (15.1)$$

This allows explicit formulas for the local and global errors and provides much new insight.

Applying a Runge-Kutta method to (15.1) yields

$$\begin{aligned} g_i &= y_0 + h \sum_{j=1}^s a_{ij} \left(\lambda(g_j - \varphi(x_0 + c_j h)) + \varphi'(x_0 + c_j h) \right) \\ y_1 &= y_0 + h \sum_{j=1}^s b_j \left(\lambda(g_j - \varphi(x_0 + c_j h)) + \varphi'(x_0 + c_j h) \right). \end{aligned} \quad (15.2)$$

If we replace here the g_i, y_0 and y_1 by the exact solution values $\varphi(x_0 + c_i h)$, $\varphi(x_0)$ and $\varphi(x_0 + h)$, respectively, we obtain a defect which is given by

$$\begin{aligned} \varphi(x_0 + c_i h) &= \varphi(x_0) + h \sum_{j=1}^s a_{ij} \varphi'(x_0 + c_j h) + \Delta_{i,h}(x_0) \\ \varphi(x_0 + h) &= \varphi(x_0) + h \sum_{j=1}^s b_j \varphi'(x_0 + c_j h) + \Delta_{0,h}(x_0). \end{aligned} \quad (15.3)$$

Taylor series expansion of the functions in (15.3) shows that

$$\Delta_{0,h}(x_0) = \mathcal{O}(h^{p+1}), \quad \Delta_{i,h}(x_0) = \mathcal{O}(h^{q+1}), \quad (15.4)$$

where p is the order of the quadrature formula (b_i, c_i) and q is the largest number such that the condition $C(q)$ (see Sect. IV.5), i.e.,

$$\sum_{j=1}^s a_{ij} c_j^{k-1} = \frac{c_i^k}{k} \quad \text{for } k = 1, \dots, q \quad \text{and all } i, \quad (15.5)$$

holds. The minimum of q and p is often called the *stage order* of the Runge-Kutta method. Subtracting (15.3) from (15.2) and eliminating the internal stages we get

$$y_1 - \varphi(x_0 + h) = R(z)(y_0 - \varphi(x_0)) - zb^T(I - zA)^{-1} \Delta_h(x_0) - \Delta_{0,h}(x_0) \quad (15.6)$$

where we have used the notation $z = \lambda h$, $R(z) = 1 + zb^T(I - zA)^{-1} \mathbb{1}$ for the stability function and $\Delta_h(x) = (\Delta_{1,h}(x), \dots, \Delta_{s,h}(x))^T$. We also denote the *local error*, which we get from (15.6) on putting $y_0 = \varphi(x_0)$, by

$$\delta_h(x) = -zb^T(I - zA)^{-1} \Delta_h(x) - \Delta_{0,h}(x). \quad (15.7)$$

If we repeat the above calculation with x_n instead of x_0 we obtain the recursion

$$y_{n+1} - \varphi(x_{n+1}) = R(z)(y_n - \varphi(x_n)) + \delta_h(x_n) \quad (15.8)$$

which leads to the following formula for the *global error*

$$y_{n+1} - \varphi(x_{n+1}) = R(z)^{n+1}(y_0 - \varphi(x_0)) + \sum_{j=0}^n R(z)^{n-j} \delta_h(x_j). \quad (15.9)$$

The classical (non-stiff) theory treats the case where $z = \mathcal{O}(h)$ and in this situation the global error behaves like $\mathcal{O}(h^p)$. When solving stiff differential equations one is interested in step sizes h which are much larger than $|\lambda|^{-1}$. We therefore study the global error (15.9) under the assumption that simultaneously $h \rightarrow 0$ and $z = \lambda h \rightarrow \infty$. In Table 15.1 we collect the results for the Runge-Kutta methods of Sect. IV.5. There in the last column (variable h) the symbols h and z have to be interpreted as $\max h_i$ and $z = \lambda \min h_i$. We remark that Formulas (15.7) and (15.8) (but not (15.9)) remain valid for variable h , if z is replaced by $z_n = h_n \lambda$.

Table 15.1. Error for (15.1) when $h \rightarrow 0$ and $z = h\lambda \rightarrow \infty$

Method		local error	global error	
			constant h	variable h
Gauss	$\begin{cases} s & \text{odd} \\ s & \text{even} \end{cases}$	h^{s+1}	$\begin{cases} h^{s+1} \\ h^s \end{cases}$	h^s
Radau IA		h^s	h^s	h^s
Radau IIA		$z^{-1} h^{s+1}$	$z^{-1} h^{s+1}$	$z^{-1} h^{s+1}$
Lobatto IIIA	$\begin{cases} s & \text{odd} \\ s & \text{even} \end{cases}$	$z^{-1} h^{s+1}$	$\begin{cases} z^{-1} h^s \\ z^{-1} h^{s+1} \end{cases}$	$z^{-1} h^s$
Lobatto IIIB	$\begin{cases} s & \text{odd} \\ s & \text{even} \end{cases}$	zh^{s-1}	$\begin{cases} zh^{s-2} \\ zh^{s-1} \end{cases}$	zh^{s-2}
Lobatto IIIC		$z^{-1} h^s$	$z^{-1} h^s$	$z^{-1} h^s$

Verification of Table 15.1.

Gauss. Since the Runge-Kutta matrix A is invertible, we have $-zb^T(I - zA)^{-1} = b^T A^{-1} + \mathcal{O}(z^{-1})$ and (15.4) inserted into (15.7) gives $\delta_h(x) = \mathcal{O}(h^{s+1})$ (observe that $q = s$). It then follows from (15.8) (for constant and variable h) that the global error behaves like $\mathcal{O}(h^s)$ because $|R(z)| \leq 1$. For odd s we have $R(\infty) = -1$ and the global error estimate can be improved in the case of constant step sizes. This follows from partial summation

$$\sum_{j=0}^n \varrho^{n-j} \delta(x_j) = \frac{1 - \varrho^{n+1}}{1 - \varrho} \delta(x_0) + \sum_{j=1}^n \frac{1 - \varrho^{n+1-j}}{1 - \varrho} (\delta(x_j) - \delta(x_{j-1})) \quad (15.10)$$

of the sum in (15.9) and from the fact that $\delta_h(x_j) - \delta_h(x_{j-1}) = \mathcal{O}(h^{q+2})$.

Radau IA. The local error estimate follows in the same way as for the Gauss methods. Since $R(z) = \mathcal{O}(z^{-1})$ the error propagation in (15.8) is negligible and the local and global errors have the same asymptotic behaviour.

Radau IIA and Lobatto IIIC. These methods have $a_{si} = b_i$ for all i . Therefore the last internal stage is identical to the numerical solution and the local error can be written as

$$\delta_h(x) = -e_s^T (I - zA)^{-1} \Delta_h(x).$$

Since A is invertible this formula shows the presence of z^{-1} in the local error. Again we have $R(\infty) = 0$, so that the global error is essentially equal to the local error.

Lobatto IIIA. The first stage is explicit, $g_1 = y_0$, and is done without introducing an error. Therefore $\Delta_{1,h}(x) = 0$ and (because of $a_{si} = b_i$) the local error has the form

$$\delta_h(x) = -e_{s-1}^T (I - z\tilde{A})^{-1} \tilde{\Delta}_h(x)$$

where $\tilde{A} = (a_{ij})_{i,j=2}^s$ and $\tilde{\Delta}_h = (\Delta_{2,h}, \dots, \Delta_{s,h})^T$. The statements of Table 15.1 now follow as for the Gauss methods.

Lobatto IIIB. The matrix A is singular (its last column vanishes), therefore the two “ z ” in (15.7) do not simply cancel for $z \rightarrow \infty$. A more detailed analysis (see Exercise 5 below) shows that the local error is not bounded if $z \rightarrow \infty$. Although A -stable, these methods are not suited for the solution of stiff problems. \square

We observe from Table 15.1 that the order of convergence for problem (15.1) with large λ is considerably smaller than the classical order. Further we see that methods satisfying $a_{si} = b_i$ (Radau IIA, Lobatto IIIA and Lobatto IIIC) give an asymptotically exact result for $z \rightarrow \infty$. Prothero & Robinson (1974) call such methods *stiffly accurate*. The importance of this condition will appear again when we treat singularly perturbed and differential-algebraic problems (Chapter VI).

The Local Error

Das besondere Schmerzenskind sind die Fehlerabschätzungen.
(L. Collatz 1950)

Our next aim is to extend the above results to general nonlinear differential equations $y' = f(x, y)$ satisfying a one-sided Lipschitz condition

$$\langle f(x, y) - f(x, z), y - z \rangle \leq \nu \|y - z\|^2. \quad (15.11)$$

The following analysis, begun by Frank, Schneid & Ueberhuber (1981), was elaborated by Frank, Schneid & Ueberhuber (1985) and Dekker & Verwer (1984). We again denote the local error by

$$\delta_h(x) = y_1 - y(x + h),$$

where y_1 is the numerical solution with initial value $y_0 = y(x)$ on the exact solution.

Proposition 15.1. *Consider a differential equation which satisfies (15.11). Assume that the Runge-Kutta matrix A is invertible, $\alpha_0(A^{-1}) \geq 0$ (see Definition 14.1), and that the stage order is q .*

a) *If $\alpha_0(A^{-1}) > 0$ then*

$$\|\delta_h(x)\| \leq C h^{q+1} \max_{\xi \in [x, x+h]} \|y^{(q+1)}(\xi)\| \quad \text{for } h\nu \leq \alpha < \alpha_0(A^{-1}).$$

b) *If $\alpha_D(A^{-1}) = 0$ for some positive diagonal matrix D and $\nu < 0$ then*

$$\|\delta_h(x)\| \leq C \left(h + \frac{1}{|\nu|} \right) h^q \max_{\xi \in [x, x+h]} \|y^{(q+1)}(\xi)\|.$$

In both cases the constant C depends only on the coefficients of the Runge-Kutta matrix and on α (for case (a)).

Remarks. a) The crucial fact in these estimates is that the right-hand side depends only on derivatives of the exact solution and not on the stiffness of the problem. These estimates are useful when a “smooth” solution of a stiff problem has to be approximated.

b) The hypothesis $\alpha_D(A^{-1}) = 0$ (see case (b)) is stronger than $\alpha_0(A^{-1}) = 0$ (see Exercise 4 below). For the Lobatto IIIC methods, for which $\alpha_0(A^{-1}) = 0$ ($s > 2$), we have $\alpha_D(A^{-1}) = 0$ with $D = B$ (see (14.21)). For stiffly accurate methods the estimate of part (b) can be improved by using (14.12) instead of (14.13).

c) In the estimates of the above proposition the maximum is taken over $\xi \in [x, x + h]$. In the case where $0 \leq c_i \leq 1$ is not satisfied, this interval must of course be correspondingly enlarged.

Proof. We put $\widehat{g}_i = y(x_0 + c_i h)$, so that the relation (14.11a) is satisfied with

$$\delta_i = y(x_0 + c_i h) - y(x_0) - h \sum_{j=1}^s a_{ij} y'(x_0 + c_j h).$$

Taylor expansion shows that

$$\|\delta_i\| \leq C_i h^{q+1} \max_{x \in [x_0, x_1]} \|y^{(q+1)}(x)\|$$

where $C_i = (|c_i|^{q+1} + (q+1) \sum_{j=1}^s |a_{ij}| \cdot |c_j|^q) / (q+1)!$ is a method-dependent constant. Similarly, the value \widehat{y}_1 of (14.11b) satisfies

$$y(x_0 + h) - \widehat{y}_1 = y(x_0 + h) - y(x_0) - h \sum_{j=1}^s b_j y'(x_0 + c_j h) = \mathcal{O}(h^{q+1}), \quad (15.12)$$

because the order of the quadrature formula (b_i, c_i) is $\geq q$. Since

$$\|\delta_h(x)\| \leq \|y_1 - \widehat{y}_1\| + \|\widehat{y}_1 - y(x_0 + h)\|$$

the desired estimates follow from (14.13) of Theorem 14.3. \square

Error Propagation

At the end of Sect. IV.12 we derived for some particular Runge-Kutta methods sharp estimates of the form

$$\|\widehat{y}_1 - y_1\| \leq \varphi_B(h\nu) \|\widehat{y}_0 - y_0\|, \quad (15.13)$$

where \widehat{y}_1, y_1 are the numerical solutions corresponding to \widehat{y}_0, y_0 , respectively, and where the differential equation satisfies (15.11). We give here a simple proof of a crude estimate of $\varphi_B(h\nu)$ which, however, will be sufficient to derive interesting convergence results.

Proposition 15.2 (Dekker & Verwer 1984). *Suppose that the differential equation satisfies (15.11) and apply an algebraically stable Runge-Kutta method with invertible A and $\alpha_0(A^{-1}) > 0$. Then for any α with $0 < \alpha < \alpha_0(A^{-1})$ there exists a constant $C > 0$ such that*

$$\|\widehat{y}_1 - y_1\| \leq (1 + Ch\nu) \|\widehat{y}_0 - y_0\| \quad \text{for } 0 \leq h\nu \leq \alpha.$$

Proof. From (12.7) we have (using the notation of the proof of Theorem 12.4)

$$\|\Delta y_1\|^2 = \|\Delta y_0\|^2 + 2 \sum_{i=1}^s b_i \langle \Delta f_i, \Delta g_i \rangle - \sum_{i=1}^s \sum_{j=1}^s m_{ij} \langle \Delta f_i, \Delta f_j \rangle. \quad (15.14)$$

By algebraic stability the last term in (15.14) is non-positive and can be neglected. Using (15.11) and the estimate (14.12) with $\delta_i = \widehat{y}_0 - y_0$ we obtain

$$\begin{aligned} 2 \sum_{i=1}^s b_i \langle \Delta f_i, \Delta g_i \rangle &\leq 2h\nu \sum_{i=1}^s b_i \|\Delta g_i\|^2 \\ &\leq 2h\nu C_1 \|\Delta g\|_D^2 \leq \frac{2h\nu C_2}{(\alpha_D(A^{-1}) - h\nu)^2} \|\Delta y_0\|^2. \end{aligned}$$

Inserting this into (15.14) yields

$$\|\Delta y_1\| \leq \left(1 + \frac{h\nu C_2}{(\alpha_D(A^{-1}) - h\nu)^2} \right) \|\Delta y_0\|$$

which proves the desired estimate. \square

B-Convergence for Variable Step Sizes

We are now in a position to present the main result of this section.

Theorem 15.3. *Consider an algebraically stable Runge-Kutta method with invertible A and stage order $q \leq p$ and suppose that (15.11) holds.*

a) If $0 < \alpha < \alpha_0(A^{-1})$ and $\nu > 0$ then the global error satisfies

$$\|y_n - y(x_n)\| \leq h^q \frac{(e^{C_1\nu(x_n-x_0)} - 1)}{C_1\nu} C_2 \max_{x \in [x_0, x_n]} \|y^{(q+1)}(x)\| \quad \text{for } h\nu \leq \alpha.$$

b) If $\alpha_0(A^{-1}) > 0$ and $\nu \leq 0$ then

$$\|y_n - y(x_n)\| \leq h^q (x_n - x_0) C_2 \max_{x \in [x_0, x_n]} \|y^{(q+1)}(x)\| \quad \text{for all } h > 0.$$

c) If $\alpha_D(A^{-1}) = 0$ for some positive diagonal matrix D and $\nu < 0$ then

$$\|y_n - y(x_n)\| \leq h^{q-1} C \left(h + \frac{1}{|\nu|} \right) (x_n - x_0) \max_{x \in [x_0, x_n]} \|y^{(q+1)}(x)\|.$$

The constants C_1, C_2, C depend only on the coefficients of the Runge-Kutta matrix. In the case of variable step sizes, h has to be interpreted as $h = \max h_i$.

Proof. This convergence result is obtained in exactly the same way as that for non-stiff problems (Theorem II.3.6). For the transported errors E_j (see Fig. II.3.2) we have the estimate (for $\nu \geq 0$)

$$\|E_j\| \leq e^{C\nu(x_n-x_j)} \|\delta_{h_{j-1}}(x_{j-1})\| \quad (15.15)$$

by Proposition 15.2, because $1 + Ch\nu \leq e^{C\nu h}$. We next insert the local error estimate of Proposition 15.1 into (15.15) and sum up the transported errors E_j . This

yields the desired estimate for $\nu \geq 0$ because

$$\begin{aligned} \sum_{j=1}^n h_{j-1} e^{C\nu(x_n-x_j)} &\leq \int_{x_0}^{x_n} e^{C\nu(x_n-x)} dx \\ &= \begin{cases} (e^{C\nu(x_n-x_0)} - 1)/(C\nu) & \text{for } \nu > 0 \\ x_n - x_0 & \text{for } \nu = 0. \end{cases} \end{aligned}$$

If $\nu < 0$ we have $\|E_j\| \leq \|\delta_{h_{j-1}}(x_{j-1})\|$ by algebraic stability and the same arguments apply. \square

Motivated by this result we define the order of B -convergence as follows:

Definition 15.4 (Frank, Schneid & Ueberhuber 1981). A Runge-Kutta method is called B -convergent of order r for problems $y' = f(x, y)$ satisfying (15.11), if the global error admits an estimate

$$\|y_n - y(x_n)\| \leq h^r \gamma(x_n - x_0, \nu) \max_{j=1, \dots, l} \max_{x \in [x_0, x_n]} \|y^{(j)}(x)\| \quad \text{for } h\nu \leq \alpha, \quad (15.16)$$

where $h = \max h_i$. Here γ is a method-dependent function and α also depends only on the coefficients of the method.

As an application of the above theorem we have

Theorem 15.5. *The Gauss and Radau IIA methods are B -convergent of order s (number of stages). The Radau IA methods are B -convergent of order $s - 1$. The 2-stage Lobatto IIIC method is B -convergent of order 1.* \square

For the Lobatto IIIC methods with $s \geq 3$ stages ($\alpha_0(A^{-1}) = 0$ and $q = s - 1$) Theorem 15.3 shows B -convergence of order $s - 2$ if $\nu < 0$. This is not an optimal result. Spijker (1986) proved B -convergence of order $s - 3/2$ for $\nu < 0$ and constant step sizes. Schneid (1987) improved this result to $s - 1$. Recently, Dekker, Kraaijevanger & Schneid (1991) showed that these methods are B -convergent of order $s - 1$ for general step size sequences, if one allows the function γ in Definition 15.4 to depend also on the ratio $\max h_i / \min h_i$.

The Lobatto IIIA and IIIB methods cannot be B -convergent since they are not algebraically stable. This will be the content of the next subsection.

***B*-Convergence Implies Algebraic Stability**

In order to find necessary conditions for *B*-convergence we consider the problem

$$y' = \lambda(x)(y - \varphi(x)) + \varphi'(x), \quad \operatorname{Re} \lambda(x) \leq \nu \quad (15.17)$$

with exact solution $\varphi(x) = x^{q+1}$. We apply a Runge-Kutta method with stage order q and obtain for the global error $\varepsilon_n = y_n - \varphi(x_n)$ the simple recursion

$$\varepsilon_{n+1} = K(Z_n)\varepsilon_n - L(Z_n)h^{q+1} \quad (15.18)$$

(cf. Eq. (15.8) of the beginning of this section, where the case $\lambda(x) = \lambda$ was treated). Here $Z_n = \operatorname{diag}(h\lambda(x_n + c_1h), \dots, h\lambda(x_n + c_sh))$ and

$$K(Z) = 1 + b^T Z(I - AZ)^{-1} \mathbb{1}, \quad L(Z) = d_0 + b^T Z(I - AZ)^{-1} d. \quad (15.19)$$

The function $K(Z)$ was already encountered in Definition 12.10, when treating *AN*-stability. The vector $d = (d_1, \dots, d_s)^T$ and d_0 in $L(Z)$ characterize the local error and are given by

$$d_0 = 1 - (q+1) \sum_{j=1}^s b_j c_j^q, \quad d_i = c_i^{q+1} - (q+1) \sum_{j=1}^s a_{ij} c_j^q. \quad (15.20)$$

Observe that by definition of the stage order we have either $d_0 \neq 0$ or $d \neq 0$ (or both). We are now in the position to prove

Theorem 15.6 (Dekker, Kraaijevanger & Schneid 1991). *Consider a DJ-irreducible Runge-Kutta method which satisfies $0 \leq c_1 < c_2 < \dots < c_s \leq 1$. If, for some r , l and $\nu < 0$, the global error satisfies the *B*-convergence estimate (15.16), then the method is algebraically stable.*

Proof. Suppose that the method is not algebraically stable. Then, by Theorem 12.13 and Lemma 15.17 below, there exists $Z = \operatorname{diag}(z_1, \dots, z_s)$ with $\operatorname{Re} z_j < 0$ such that $(I - AZ)^{-1}$ exists and

$$|K(Z)| > 1, \quad L(Z) \neq 0. \quad (15.21)$$

We consider the interval $[0, (1 + \theta)/2]$ and for even N the step size sequence $(h_n)_{n=0}^{N-1}$ given by

$$h_n = 1/N \quad (\text{for } n \text{ even}), \quad h_n = \theta/N \quad (\text{for } n \text{ odd}).$$

If N is sufficiently large it is possible to define a function $\lambda(x)$ which satisfies $\operatorname{Re} \lambda(x) \leq \nu$ and

$$\lambda(x_n + c_i h_n) = \begin{cases} Nz_i & \text{for } n \text{ even} \\ Nz_{s+1-i} & \text{for } n \text{ odd} \end{cases}.$$

Because of (15.18) the global error $\varepsilon_n = y_n - \varphi(x_n)$ for the problem (15.17) sat-

isfies (with $h = 1/N$)

$$\begin{aligned}\varepsilon_{n+1} &= K(Z)\varepsilon_n - h^{q+1}L(Z) \quad \text{for } n \text{ even} \\ \varepsilon_{n+1} &= K(\tilde{Z})\varepsilon_n - h^{q+1}L(\tilde{Z}) \quad \text{for } n \text{ odd}\end{aligned}$$

where $\tilde{Z} = \text{diag}(\theta z_s, \dots, \theta z_1)$. Consequently we have

$$\varepsilon_{2m+2} = K(\tilde{Z})K(Z)\varepsilon_{2m} - h^{q+1}(K(\tilde{Z})L(Z) + \theta^{q+1}L(\tilde{Z}))$$

and the error at $X = (1 + \theta)/2$ is given by

$$\varepsilon_N = -\frac{1}{N^{q+1}} (K(\tilde{Z})L(Z) + \theta^{q+1}L(\tilde{Z})) \frac{(K(\tilde{Z})K(Z))^{N/2} - 1}{K(\tilde{Z})K(Z) - 1}. \quad (15.22)$$

If θ is sufficiently small, $K(\tilde{Z}) \rightarrow 1$ and $L(\tilde{Z}) \rightarrow d_0$, so that by (15.21)

$$|K(\tilde{Z})K(Z)| > 1 \quad \text{and} \quad K(\tilde{Z})L(Z) + \theta^{q+1}L(\tilde{Z}) \neq 0.$$

Therefore $|\varepsilon_N| \rightarrow \infty$ as $N \rightarrow \infty$ (N even), which contradicts the estimate (15.16) of B -convergence. \square

To complete the above proof we give the following lemma:

Lemma 15.7 (Dekker, Kraaijevanger & Schneid 1990). *Consider a DJ -irreducible Runge-Kutta method and suppose*

$$b^T Z(I - AZ)^{-1}d = 0 \quad (15.23)$$

for all $Z = \text{diag}(z_1, \dots, z_s)$ with $I - AZ$ invertible; then $d = 0$.

Proof. We define

$$T = \{j \mid b_{i_1} a_{i_1 i_2} a_{i_2 i_3} \dots a_{i_{k-1} i_k} = 0 \quad \text{for all } k \text{ and } i_l \text{ with } i_k = j\}.$$

Putting $k = 1$ we obtain $b_j = 0$ for $j \in T$. Further, if $i \notin T$ and $j \in T$ there exists (i_1, \dots, i_k) with $i_k = i$ such that

$$b_{i_1} a_{i_1 i_2} \dots a_{i_{k-1} i_k} \neq 0, \quad b_{i_1} a_{i_1 i_2} \dots a_{i_{k-1} i_k} a_{i_j} = 0$$

implying $a_{ij} = 0$. Therefore the method is DJ -reducible if $T \neq \emptyset$. For the proof of the statement it thus suffices to show that $d \neq 0$ implies $T \neq \emptyset$.

Replacing $(I - AZ)^{-1}$ by its geometric series, assumption (15.23) becomes equivalent to

$$b^T Z(AZ)^{k-1}d = 0 \quad \text{for all } k \text{ and } Z = \text{diag}(z_1, \dots, z_s). \quad (15.24)$$

Comparing the coefficient of $z_{i_1} \dots z_{i_k}$ gives

$$\sum b_{j_1} a_{j_1 j_2} \dots a_{j_{k-1} j_k} d_{j_k} = 0, \quad (15.25)$$

where the summation is over all permutations (j_1, \dots, j_k) of (i_1, \dots, i_k) . Suppose now that $d_j \neq 0$ for some index j . We shall prove by induction on k that

$$b_{i_1} a_{i_1 i_2} \dots a_{i_{k-1} i_k} = 0 \quad \text{for all } i_\ell \ (\ell = 1, \dots, k) \quad \text{with } i_k = j, \quad (15.26)$$

so that $j \in T$ and consequently $T \neq \emptyset$.

For $k = 1$ this follows immediately from (15.25). In order to prove (15.26) for $k + 1$ we suppose, by contradiction, that (i_1, \dots, i_{k+1}) with $i_{k+1} = j$ exists such that $b_{i_1} a_{i_1 i_2} \dots a_{i_k i_{k+1}} \neq 0$. The relation (15.25) then implies the existence of a permutation (j_1, \dots, j_{k+1}) of (i_1, \dots, i_{k+1}) such that $b_{j_1} a_{j_1 j_2} \dots a_{j_k j_{k+1}} \neq 0$, too. We now denote by q the smallest index for which $i_q \neq j_q$. Then $i_q = j_r$ for some $r > q$ and

$$b_{i_1} a_{i_1 i_2} \dots a_{i_{q-1} i_q} a_{j_r j_{r+1}} \dots a_{j_k j_{k+1}} \neq 0 \quad (15.27)$$

contradicts the induction hypothesis, because the expression in (15.27) contains at most k factors. \square

The Trapezoidal Rule

The trapezoidal rule

$$y_{k+1} = y_k + \frac{h_k}{2} (f(x_k, y_k) + f(x_{k+1}, y_{k+1})) \quad (15.28)$$

is not algebraically stable. Therefore (Theorem 15.6) it cannot be B -convergent in the sense of Definition 15.4. Nevertheless it is possible to derive estimates (15.16), if we restrict ourselves to special step size sequences (constant, monotonic, ...). This was first proved by Stetter (unpublished) and investigated in detail by Kraaijevanger (1985). The result is

Theorem 15.8 (Kraaijevanger 1985). *If the differential equation satisfies (15.11), then the global error of the trapezoidal rule permits for $h_j \nu \leq \alpha < 2$ the estimate*

$$\|y_n - y(x_n)\| \leq C \max_{x \in [x_0, x_n]} \|y^{(3)}(x)\| \sum_{k=0}^{n-1} \left(\prod_{j=k+1}^{n-1} \max(1, h_j/h_{j-1}) \right) h_k^3.$$

Proof. We denote by $\hat{y}_k = y(x_k)$ the exact solution at the grid points. From the Taylor expansion we then get

$$\hat{y}_{k+1} = \hat{y}_k + \frac{h_k}{2} (f(x_k, \hat{y}_k) + f(x_{k+1}, \hat{y}_{k+1})) + \delta_k \quad (15.29)$$

where

$$\|\delta_k\| \leq \frac{1}{12} h_k^3 \max_{x \in [x_k, x_{k+1}]} \|y^{(3)}(x)\|. \quad (15.30)$$

The main idea is now to introduce the intermediate values

$$\begin{aligned} y_{k+1/2} &= y_k + \frac{h_k}{2} f(x_k, y_k) = y_{k+1} - \frac{h_k}{2} f(x_{k+1}, y_{k+1}) \\ \widehat{y}_{k+1/2} &= \widehat{y}_k + \frac{h_k}{2} f(x_k, \widehat{y}_k) + \delta_k = \widehat{y}_{k+1} - \frac{h_k}{2} f(x_{k+1}, \widehat{y}_{k+1}). \end{aligned} \quad (15.31)$$

The transition $y_{k-1/2} \rightarrow y_{k+1/2}$

$$y_{k+1/2} = y_{k-1/2} + \frac{1}{2}(h_{k-1} + h_k)f(x_k, y_k)$$

can then be interpreted as one step of the θ -method

$$y_{m+1} = y_m + hf(x_m + \theta h, y_m + \theta(y_{m+1} - y_m))$$

with $\theta = h_{k-1}/(h_{k-1} + h_k)$ and step size $h = (h_{k-1} + h_k)/2$. A similar calculation shows that the same θ -method maps $\widehat{y}_{k-1/2}$ to $\widehat{y}_{k+1/2} - \delta_k$. Therefore we have

$$\|\widehat{y}_{k+1/2} - y_{k+1/2} - \delta_k\| \leq \varphi_B(h\nu) \|\widehat{y}_{k-1/2} - y_{k-1/2}\|,$$

where the growth function $\varphi_B(h\nu)$ is given by (see Eqs. (12.42) and (11.13))

$$\begin{aligned} \varphi_B(h\nu) &= \max\{(1 - \theta)/\theta, (1 + (1 - \theta)h\nu)/(1 - \theta h\nu)\} \\ &= \max\{h_k/h_{k-1}, (1 + \frac{1}{2}h_k\nu)/(1 - \frac{1}{2}h_{k-1}\nu)\} =: \varphi_k. \end{aligned} \quad (15.32)$$

By the triangle inequality we also get

$$\|\widehat{y}_{k+1/2} - y_{k+1/2}\| \leq \varphi_k \|\widehat{y}_{k-1/2} - y_{k-1/2}\| + \|\delta_k\|. \quad (15.33)$$

Further it follows from (15.31) with $k = 0$ and from $\widehat{y}_0 = y_0$ that

$$\|\widehat{y}_{1/2} - y_{1/2}\| = \|\delta_0\|, \quad (15.34)$$

whereas the backward Euler steps $y_{n-1/2} \rightarrow y_n$ and $\widehat{y}_{n-1/2} \rightarrow \widehat{y}_n$ (see (15.31)) imply

$$\|\widehat{y}_n - y_n\| \leq \frac{1}{(1 - \frac{1}{2}h_{n-1}\nu)} \|\widehat{y}_{n-1/2} - y_{n-1/2}\| \quad (15.35)$$

again by Example 12.24 with $\theta = 1$. A combination of (15.33), (15.34) and (15.35) yields

$$\|\widehat{y}_n - y_n\| \leq \frac{1}{(1 - \frac{1}{2}h_{n-1}\nu)} \sum_{k=0}^{n-1} \left(\prod_{j=k+1}^{n-1} \varphi_j \right) \|\delta_k\|. \quad (15.36)$$

For $\nu \leq 0$ we have $\varphi_k \leq \max(1, h_k/h_{k-1})$ and the statement follows if we insert (15.30) into (15.36). For $\nu \geq 0$ we use the estimate $(h_{k-1}\nu \leq 1)$

$$\frac{1 + \frac{1}{2}h_k\nu}{1 - \frac{1}{2}h_{k-1}\nu} = \frac{1 + \frac{1}{2}h_{k-1}\nu}{1 - \frac{1}{2}h_{k-1}\nu} \cdot \frac{1 + \frac{1}{2}h_k\nu}{1 + \frac{1}{2}h_{k-1}\nu} \leq e^{2h_{k-1}\nu} \cdot \max\left(1, \frac{h_k}{h_{k-1}}\right)$$

so that the statement holds with $C = e^{2\nu(x_n - x_0)}/12$. \square

Corollary 15.9. If the step size sequence $(h_k)_{k=0}^{N-1}$ is constant or monotonic, then for $h = \max h_i$

$$\|y_n - y(x_n)\| \leq C \max_{x \in [x_0, x_n]} \|y^{(3)}(x)\| \cdot h^2. \quad \square$$

Order Reduction for Rosenbrock Methods

Obviously, Rosenbrock methods (Definition 7.1) cannot be B -convergent in the sense of Definition 15.4 (see also Exercise 7 of Sect. IV.12). Nevertheless it is interesting to study their behaviour on stiff problems such as the Prothero & Robinson model (15.1). Since this equation is non-autonomous we have to use the formulation (7.4a). A straightforward calculation shows that the global error $\varepsilon_n = y_n - \varphi(x_n)$ satisfies the recursion

$$\varepsilon_{n+1} = R(z)\varepsilon_n + \delta_h(x_n) \quad (15.37)$$

where $R(z)$ is the stability function (7.14) and the local error is given by

$$\delta_h(x) = \varphi(x) - \varphi(x+h) + b^T(I - zB)^{-1}\Delta \quad (15.38)$$

with $B = (\alpha_{ij} + \gamma_{ij})$, $b = (b_1, \dots, b_s)^T$, $\Delta = (\Delta_1, \dots, \Delta_s)^T$ and

$$\Delta_i = z(\varphi(x) - \varphi(x + \alpha_i h) - \gamma_i h \varphi'(x)) + h \varphi'(x + \alpha_i h) + \gamma_i h^2 \varphi''(x).$$

Taylor expansion gives the following result.

Lemma 15.10. *The local error $\delta_h(x)$ of a Rosenbrock method applied to (15.1) satisfies for $h \rightarrow 0$ and $z = \lambda h \rightarrow \infty$*

$$\delta_h(x) = \left(\sum_{i,j} b_i \omega_{ij} \alpha_j^2 - 1 \right) \frac{h^2}{2} \varphi''(x) + \mathcal{O}(h^3) + \mathcal{O}\left(\frac{h^2}{z}\right),$$

where ω_{ij} are the entries of B^{-1} . \square

Remarks. a) Unless the Rosenbrock method satisfies the new order condition

$$\sum_{i,j=1}^s b_i \omega_{ij} \alpha_j^2 = 1, \quad (15.39)$$

the local error and the global error (if $|R(\infty)| < 1$) are of size $\mathcal{O}(h^2)$. Since none of the classical Rosenbrock methods of Sect. IV.7 satisfies (15.39), their order of convergence is only 2 for the problem (15.1) if λ is very large.

b) A convenient way to satisfy (15.39) is to require

$$\alpha_{si} + \gamma_{si} = b_i \quad (i = 1, \dots, s) \quad \text{and} \quad \alpha_s = 1. \quad (15.40)$$

This is the analogue of the condition $a_{si} = b_i$ for Runge-Kutta methods. It implies not only (15.39) but even

$$\delta_h(x) = \mathcal{O}\left(\frac{h^2}{z}\right),$$

so that such methods yield asymptotically exact results for $z \rightarrow \infty$.

c) A deeper understanding of Eq. (15.39) will be possible when studying the error of Rosenbrock methods for singular perturbation and differential-algebraic problems (Chapter VI). We shall construct there methods satisfying (15.40).

d) Scholz (1989) writes the local error $\delta_h(x)$ in the form

$$\delta_h(x) = \sum_{j \geq 2} C_j(z) h^j \varphi^{(j)}(x) \quad (15.41)$$

and investigates the possibility of having $C_j(z) \equiv 0$ for $j = 2$ (and also $j > 2$). Hundsdorfer (1986) and Strehmel & Weiner (1987) extend the above analysis to semi-linear problems (11.21) which satisfy (11.22). Their results are rather technical but allow the construction of “ B -convergent” methods of order $p > 1$.

Exercises

1. Prove that the stage order of an SDIRK method is at most 1, that of a DIRK method at most 2.
2. Consider a Runge-Kutta method with $0 \leq c_1 < \dots < c_s \leq 1$ which has stage order q . Prove that the method cannot be B -convergent (for variable step sizes) of order $q + 1$.

Hint. Use Formula (15.22) and prove that

$$\frac{K(\tilde{Z})L(Z) + \theta^{q+1}L(\tilde{Z})}{K(\tilde{Z})K(Z) - 1} \quad (15.42)$$

cannot be uniformly bounded for

$$Z = \text{diag}(z_1, \dots, z_s), \quad \tilde{Z} = \text{diag}(\tilde{z}_1, \dots, \tilde{z}_s)$$

with $\text{Re } z_i \leq 0$, $\text{Re } \tilde{z}_i \leq 0$ (in the case $c_1 = 0$ and $c_s = 1$ one has to prove this under the restriction $\tilde{z}_1 = \theta z_s$, $\tilde{z}_s = \theta z_1$). For this consider values z_j , \tilde{z}_j close to the origin.

3. (Burrage & Hundsdorfer 1987). Assume $c_i - c_j$ is not an integer for $1 \leq i < j \leq s$, and the order of B -convergence (for constant step sizes) of a Runge-Kutta method is $q + 1$ (q denotes the stage order). Then $d_0 = 0$ and all components of $d = (d_1, \dots, d_s)^T$ are equal (see (15.20) for the definition of d_j).

Hint. Study the uniform boundedness of the function $L(Z)/(K(Z) - 1)$.

4. (Kraaijevanger). Show that for

$$A^{-1} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \quad (15.43)$$

we have $\alpha_0(A^{-1}) = 0$, but there exists no positive diagonal matrix D such that $\alpha_D(A^{-1}) = 0$. For more insight see “Corollary 2.15” of Kraaijevanger & Schneid (1991).

5. Prove that for the Lobatto IIIB methods, with

$$A = \begin{pmatrix} \tilde{A} & 0 \\ a^T & 0 \end{pmatrix}$$

the dominant term of the local error (15.7) is (for $h \rightarrow 0$ and $z = h\lambda \rightarrow \infty$)

$$zb_s(a^T \tilde{A}^{-1} c^{q+1} - 1) \frac{h^{q+1}}{(q+1)!} \varphi^{(q+1)}(x).$$

Here $q = s - 2$ is the stage order and $c = (c_1, \dots, c_{s-1})^T$. Show further that

$$a^T \tilde{A}^{-1} c^k = 1 \quad \text{for } k = 1, 2, \dots, q \quad (15.44)$$

$$a^T \tilde{A}^{-1} c^k \neq 1 \quad \text{for } k = q + 1. \quad (15.45)$$

Hint. Equation (15.44) follows from $C(q)$. Show (15.45) by supposing that $a^T \tilde{A}^{-1} c^{q+1} = 1$ which together with (15.44) implies that

$$\sum_{i=1}^{s-1} d_i p(c_i) = p(1) \quad \text{where} \quad d^T = a^T \tilde{A}^{-1}$$

for every polynomial of $\deg p \leq q + 1 = s - 1$ satisfying $p(0) = 0$. Arrive at a contradiction with

$$p(x) = (x - c_1)(x - c_2) \cdots (x - c_{s-1}).$$

Chapter V. Multistep Methods for Stiff Problems

Multistep methods (BDF) were the first numerical methods to be proposed for stiff differential equations (Curtiss & Hirschfelder 1952) and since Gear's book (1971) computer codes based on these methods have been the most prominent and most widely used for all stiff computations.

This chapter introduces the linear stability theory for multistep methods in Sect. V.1, and arrives at the famous theorem of Dahlquist which says that A -stable multistep methods cannot have high order. Attempts to circumvent this barrier proceed mainly in two directions: either study methods with slightly weaker stability requirements (Sect. V.2) or introduce new classes of methods (Sect. V.3). Order star theory on Riemann surfaces (Sect. V.4) then helps to extend Dahlquist's barrier to generalized methods and to explain various properties of stability domains. Section V.5 presents numerical experiments with several codes based on the methods introduced.

Since all the foregoing stability theory is based uniquely on linear autonomous problems $y' = Ay$, the question arises of their validity for general nonlinear problems. This leads to the concepts of G -stability for multistep methods (Sect. V.6) and algebraic stability for general linear methods (Sect. V.9).

Another important subject is convergence estimates for $h \rightarrow 0$ which are independent of the stiffness (the analogue of B -convergence in Sect. IV.15). We describe various techniques for obtaining such estimates in Sections V.7 (for linear problems) as well as V.6 and V.8 (for nonlinear problems). These techniques are: use of G -stability, the Kreiss matrix theorem, the multiplier technique and, last but not least, a discrete variation of constants formula.

V.1 Stability of Multistep Methods

A general k -step multistep method is of the form

$$\alpha_k y_{m+k} + \alpha_{k-1} y_{m+k-1} + \dots + \alpha_0 y_m = h(\beta_k f_{m+k} + \dots + \beta_0 f_m). \quad (1.1)$$

For this method, we can do the same stability analysis as in Sect. IV.2 for Euler's method. This means that we apply it to the linearized and autonomous system

$$y' = Jy \quad (1.2)$$

(see (IV.2.2')); this gives

$$\alpha_k y_{m+k} + \dots + \alpha_0 y_m = hJ(\beta_k y_{m+k} + \dots + \beta_0 y_m). \quad (1.3)$$

We again introduce a new basis for the vectors y_{m+i} consisting of the eigenvectors of J . Then for the *coefficients* of y_{m+i} , with respect to an eigenvector v of J , we obtain exactly the same recurrence equation as (1.3), with J replaced by the corresponding eigenvalue λ . This gives ¹

$$(\alpha_k - \mu\beta_k)y_{m+k} + \dots + (\alpha_0 - \mu\beta_0)y_m = 0, \quad \mu = h\lambda \quad (1.4)$$

and is the same as Method (1.1) applied to Dahlquist's test equation

$$y' = \lambda y. \quad (1.5)$$

The Stability Region

The difference equation (1.4) is solved using Lagrange's method (see Volume I, Sect. III.3): we set $y_j = \zeta^j$, divide by ζ^m and obtain the characteristic equation

$$(\alpha_k - \mu\beta_k)\zeta^k + \dots + (\alpha_0 - \mu\beta_0) = \varrho(\zeta) - \mu\sigma(\zeta) = 0 \quad (1.6)$$

which depends on the complex parameter μ . The polynomials $\varrho(\zeta)$ and $\sigma(\zeta)$ are our old friends from (III.2.4). The difference equation (1.4) has stable solutions (for arbitrary starting values) iff all roots of (1.6) are ≤ 1 in modulus. In addition, *multiple* roots must be *strictly* smaller than 1 (see Volume I, Sect. III.3, Exercise 1).

¹ In contrast to Chapter IV, where the product $h\lambda$ was denoted throughout by z , we write $h\lambda = \mu$ here, since in multistep theory (Sect. III.3) z denotes the Cayley transform of ζ .

Definition 1.1. The set

$$S = \left\{ \mu \in \mathbb{C} ; \begin{array}{l} \text{all roots } \zeta_j(\mu) \text{ of (1.6) satisfy } |\zeta_j(\mu)| \leq 1, \\ \text{multiple roots satisfy } |\zeta_j(\mu)| < 1 \end{array} \right\} \quad (1.7)$$

is called the *stability domain* or *stability region* or *region of absolute stability* of Method (1.1). We have *A-stability* if $S \supset \mathbb{C}^-$.

It is sometimes desirable to consider S as a subset of the compactified complex plane $\overline{\mathbb{C}}$. In this case, for $\mu \rightarrow \infty$, the roots of Eq. (1.6) tend to those of $\sigma(\zeta) = 0$.

For $\mu = 0$, Eq. (1.6) becomes $\varrho(\zeta) = 0$. Thus the usual stability (in the sense of Definition III.3.2) is equivalent to $0 \in S$.

Theorem 1.2. All numerical solutions of Method (1.1) are bounded for the linearized equation (1.2) with a diagonalizable matrix J iff $h\lambda \in S$ for all eigenvalues λ of J . \square

We explain the computation of the stability domain at a particular example, the explicit Adams method of order 4 (see Sect. III.1, Eq. (1.5)),

$$y_{m+4} = y_{m+3} + h \left(\frac{55}{24} f_{m+3} - \frac{59}{24} f_{m+2} + \frac{37}{24} f_{m+1} - \frac{9}{24} f_m \right),$$

for which Eq. (1.6) becomes

$$\zeta^4 - \left(1 + \frac{55}{24} \mu \right) \zeta^3 + \frac{59}{24} \mu \zeta^2 - \frac{37}{24} \mu \zeta + \frac{9}{24} \mu = 0. \quad (1.8)$$

In Fig. 1.1 we display the complicated behavior of the roots of this equation. We choose the μ values as the dots surrounding the white horse, and plot the corresponding 4 roots $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ in the ζ -plane, which can be observed to emerge from the roots 1, 0, 0, 0 of the ϱ -polynomial.

Complex mappings are *conformal*, i.e., preserve angles and orientation. The angle of rotation and the magnification of a complex map is (locally) determined by its *derivative*. Differentiating (1.8) with respect to μ and putting $\mu = 0$, $\zeta = 1$ gives

$$\varrho'(1) \cdot \zeta_1'(0) - \sigma(1) = 0,$$

hence $\zeta_1'(0) = 1$ (because of the consistency conditions $\varrho'(1) \neq 0$, $\sigma(1) = \varrho'(1)$, see Volume I, Eq. (III.2.6)). This explains the fact that the map $\mu \mapsto \zeta_1$ is close to $1 + \mu$ in the neighbourhood of $\mu = 0$, and $\zeta_1(\mu)$ moves *inside* the unit disc when μ moves *inside* \mathbb{C}^- .

The Root Locus Curve. The key for computing S is the fact that the *inverse* map $\zeta \mapsto \mu$, since (1.8) is linear in μ , can easily be computed and is one-valued

$$\mu = \frac{\varrho(\zeta)}{\sigma(\zeta)} = \frac{\zeta^4 - \zeta^3}{\frac{55}{24} \zeta^3 - \frac{59}{24} \zeta^2 + \frac{37}{24} \zeta - \frac{9}{24}}. \quad (1.9)$$

The *outside* of the unit circle in the ζ -plane mapped back into the μ -plane by this formula (see the zodiac of gray horses in Fig. 1.1) produces the forbidden μ -values, for which at least one root $\zeta_i(\mu)$ generates instability. The image of the boundary curve of the unit circle $\zeta = e^{i\theta}$, $0 \leq \theta \leq 2\pi$, is called the *root locus curve*. It must be considered as an oriented curve and the stability region, whenever it is not empty, must lie *to the left* of it.

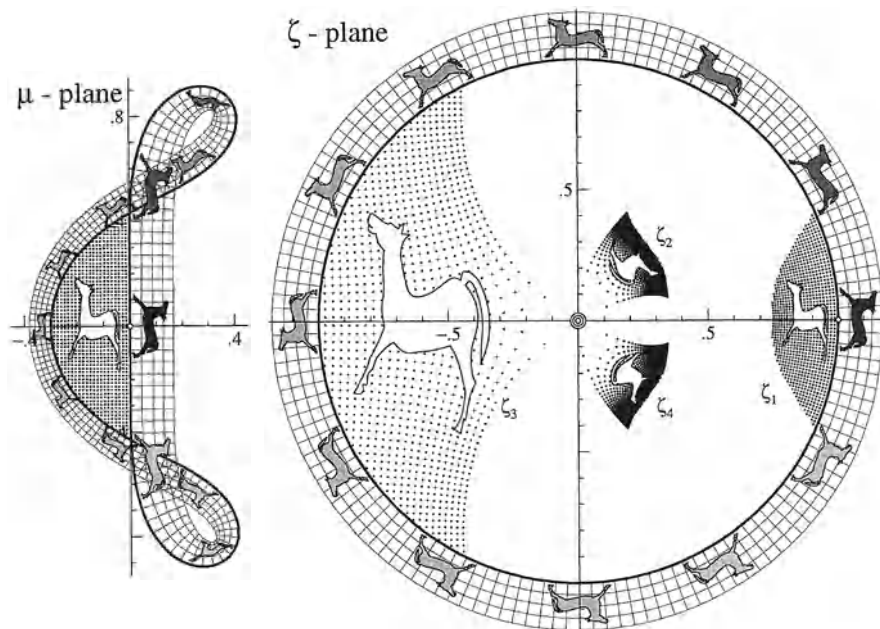


Fig. 1.1. Plot of the stability function (1.8) with root locus curve

We conclude that the stability domain of Adams4 is precisely the small diamond shaped region surrounded by the root locus curve in the positive direction located between the origin and the point $\mu = 2 \cdot 24 / (-55 - 59 - 37 - 9) = -0.3$.

Adams Methods

The *explicit Adams methods* (III.1.5) applied to $y' = \lambda y$ give

$$y_{n+1} = y_n + \mu \sum_{j=0}^{k-1} \gamma_j \nabla^j y_n, \quad \gamma_0 = 1, \gamma_1 = \frac{1}{2}, \gamma_2 = \frac{5}{12}, \gamma_3 = \frac{3}{8}, \dots \quad (1.10)$$

or, after putting $y_n = \zeta^n$ and dividing by ζ^n ,

$$\zeta - 1 = \mu \left(\gamma_0 + \gamma_1 \left(1 - \frac{1}{\zeta} \right) + \gamma_2 \left(1 - \frac{2}{\zeta} + \frac{1}{\zeta^2} \right) + \dots \right).$$

Hence the root locus curve becomes

$$\mu = \frac{\zeta - 1}{\sum_{j=0}^{k-1} \gamma_j (1 - 1/\zeta)^j}, \quad \zeta = e^{i\theta}. \quad (1.10')$$

For $k = 1$ we again obtain the circle of Euler's method, centred at -1 . These curves are plotted in Fig. 1.2 for $k = 2, 3, \dots, 6$ and show stability domains of rapidly decreasing sizes. These methods are thus surely not appropriate for stiff problems.

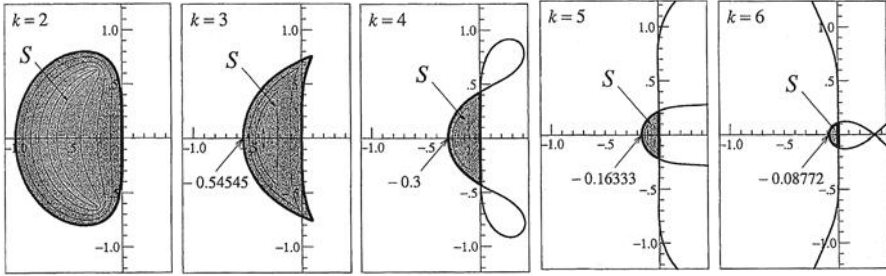


Fig. 1.2. Stability domains for explicit Adams methods

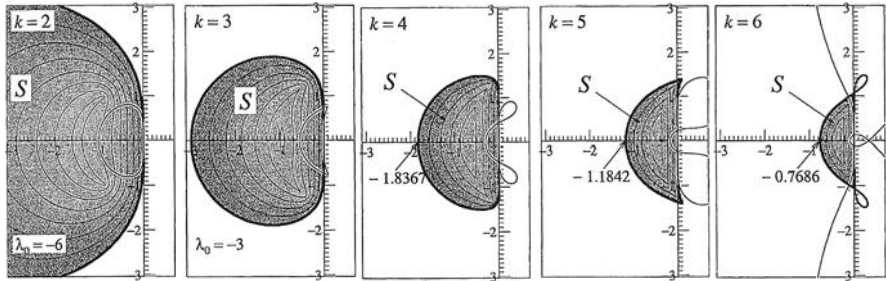


Fig. 1.3. Stability domains of implicit Adams methods, compared to those of the explicit ones

The *implicit Adams methods* (III.1.8) lead to

$$y_{n+1} = y_n + \mu \sum_{j=0}^k \gamma_j^* \nabla^j y_{n+1}, \quad \gamma_0^* = 1, \gamma_1^* = -\frac{1}{2}, \gamma_2^* = -\frac{1}{12}, \dots \quad (1.11)$$

Here we put $y_n = \zeta^n$ and divide by ζ^{n+1} . This gives

$$\mu = \frac{1 - 1/\zeta}{\sum_{j=0}^k \gamma_j^* (1 - 1/\zeta)^j}, \quad \zeta = e^{i\theta}. \quad (1.11')$$

For $k = 1$ this is the implicit trapezoidal rule and is A -stable. For $k = 2, 3, \dots, 6$ the stability domains, though much larger than those of the explicit methods, do not cover \mathbb{C}^- (see Fig. 1.3). Hence these methods are *not* A -stable.

Predictor-Corrector Schemes

The inadequacy of the theory incorporating the effect of the corrector equation only for predictor-corrector methods was first discovered through experimental computations on the prototype linear equation

$$y' = f(x, y) = -100y + 100, \quad y(0) = 0,$$

(...) Very poor correlation of actual errors with the errors expected on the basis of the properties of the corrector equation alone was obtained. This motivated the development of the theory
... (P.E. Chase 1962)

As we have seen in Sect. III.1, the classical way of computing y_{n+1} from the implicit equations (III.1.8) is to use the result y_{n+1}^* of the explicit Adams method as a *predictor* in $\beta_k f(x_{n+1}, y_{n+1})$. This destroys a good deal of the stability properties of the method (Chase 1962). The stability analysis changes as follows: the predictor formula

$$y_{n+1}^* = y_n + \mu(\gamma_0 y_n + \gamma_1(y_n - y_{n-1}) + \gamma_2(y_n - 2y_{n-1} + y_{n-2}) + \dots) \quad (1.12)$$

must be inserted into the corrector formula

$$\begin{aligned} y_{n+1} = y_n + \mu \big(& \gamma_0^* y_{n+1}^* + \\ & \gamma_1^* (y_{n+1}^* - y_n) + \\ & \gamma_2^* (y_{n+1}^* - 2y_n + y_{n-1}) + \\ & \gamma_3^* (y_{n+1}^* - 3y_n + 3y_{n-1} - y_{n-2}) + \dots \big). \end{aligned} \quad (1.13)$$

Since there is a μ in (1.12) and in (1.13), we obtain this time, by putting $y_n = \zeta^n$ and dividing by ζ^n , a *quadratic* equation for μ ,

$$A\mu^2 + B\mu + C = 0, \quad (1.14)$$

$$\begin{aligned} A &= \left(\sum_{j=0}^k \gamma_j^* \right) \left(\sum_{j=0}^{k-1} \gamma_j \left(1 - \frac{1}{\zeta} \right)^j \right), \\ B &= (1 - \zeta) \sum_{j=0}^k \gamma_j^* + \zeta \sum_{j=0}^k \gamma_j^* \left(1 - \frac{1}{\zeta} \right)^j, \\ C &= 1 - \zeta. \end{aligned}$$

For each $\zeta = e^{i\theta}$, Eq. (1.14) has two roots. These give rise to two root locus curves which determine the stability domain. These curves are represented in Fig. 1.4 and compared to those of the original implicit methods. It can be seen that we loose a lot of stability. In particular, for $k = 1$ the trapezoidal rule becomes an explicit second order Runge Kutta method and the A -stability is destroyed.

While Chase (1962) studied real eigenvalues only, the general complex case has been stated by Crane & Klopfenstein (1965) and, with beautiful figures, by

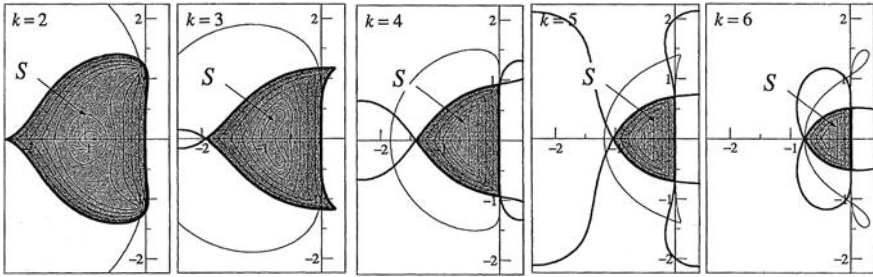


Fig. 1.4. Stability domains for PECE compared to original implicit methods

Krogh (1966). All three papers also searched for procedures with increased stability domains. This research was brought to perfection by Stetter (1968).

Nyström Methods

Thus we see that Milne's method will not handle so simple an equation as $y' = -y$, $y(0) = 1 \dots$ (R.W. Hamming 1959)

... Milne's method has a number of virtues not possessed by its principal rival, the Runge-Kutta method, which are especially important when the order of the system of equations is fairly high ($N=10$ to 30 or more) ... (R.W. Hamming 1959)

The *explicit Nyström method* (III.1.13) for $k = 1$ and 2 is the “explicit midpoint rule”

$$y_{n+1} = y_{n-1} + 2hf_n \quad (1.15)$$

and leads to the root locus curve

$$\mu = \frac{e^{i\theta} - e^{-i\theta}}{2} = i \sin \theta. \quad (1.15')$$

This curve moves up and down the imaginary axis between $\pm i$ and leaves as stability domain just the interval $(-i, +i)$ (see Fig. 1.5). All eigenvalues in the interior of the negative half plane lead to instabilities. This is caused by the second root -1 of $\varrho(\zeta)$ which moves out of the unit circle when μ goes West. This famous phenomenon is called the “weak instability” of the midpoint rule and was the “entry point” of Dahlquist's stability-career (Dahlquist 1951). The graphs of Fig. III.9.2 nicely show the (weak) instability of the numerical solution.

The *implicit Milne-Simpson method* (III.1.15) for $k = 2$ and 3 is

$$y_{n+1} = y_{n-1} + h \left(\frac{1}{3}f_{n+1} + \frac{4}{3}f_n + \frac{1}{3}f_{n-1} \right) \quad (1.16)$$

and has the root locus curve

$$\mu = \frac{e^{i\theta} - e^{-i\theta}}{\frac{1}{3}e^{i\theta} + \frac{4}{3} + \frac{1}{3}e^{-i\theta}} = 3i \frac{\sin \theta}{\cos \theta + 2}, \quad (1.16')$$

which moves up and down the imaginary axis between $\pm i\sqrt{3}$. Thus its behaviour is similar to the explicit Nyström method with a slightly larger stability interval.

The higher order Nyström and Milne-Simpson methods have root locus curves which are oriented the wrong way round (see Fig. 1.5). Their stability domains therefore reduce to the smallest possible set (for stable methods): *just the origin*.

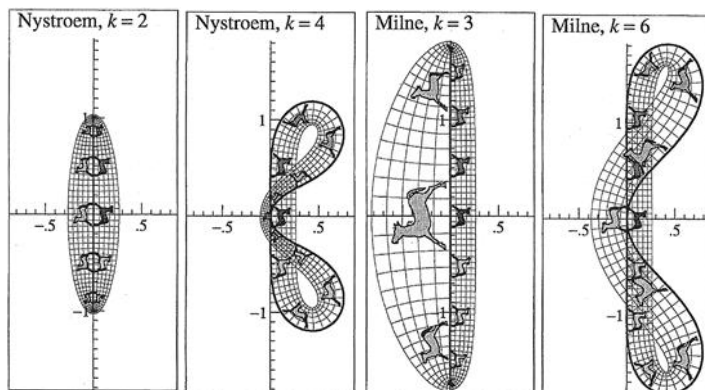


Fig. 1.5. Root locus curves for Nyström and Milne methods

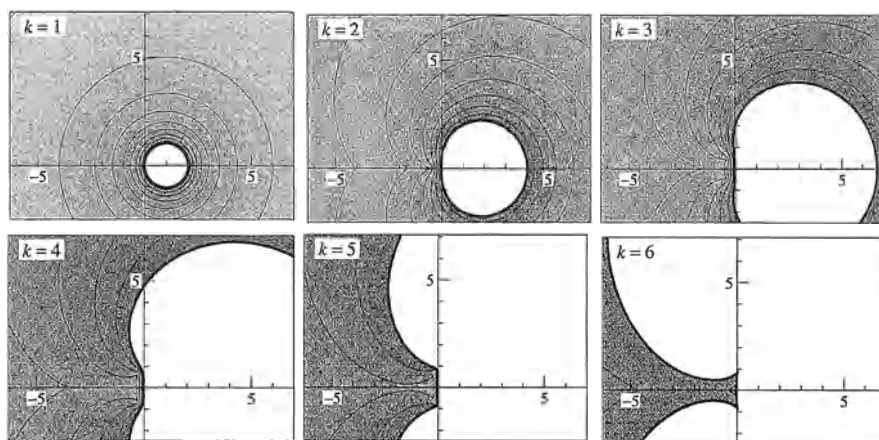


Fig. 1.6. Root locus curves and stability domains of BDF methods

BDF

The backward differentiation formulas $\sum_{j=1}^k \frac{1}{j} \nabla^j y_{n+1} = h f_{n+1}$ (see Equation (III.1.22')) have the root locus curves given by

$$\mu = \sum_{j=1}^k \frac{1}{j} \left(1 - \frac{1}{\zeta}\right)^j = \sum_{j=1}^k \frac{1}{j} (1 - e^{-i\theta})^j. \quad (1.17)$$

For $k = 1$ we have the implicit Euler method with stability domain $S = \{\mu; |\mu - 1| \geq 1\}$. For $k = 2$ the root locus curve (see Fig. 1.6) has $\operatorname{Re}(\mu) = \frac{3}{2} - 2 \cos \theta + \frac{1}{2} \cos 2\theta$ which is ≥ 0 for all θ . Therefore the method is A -stable and of order 2. However, for $k = 3, 4, 5$ and 6, we see that the methods lose more and more stability on a part of the imaginary axis. For $k \geq 7$, as we know, the formulas are unstable anyway, even at the origin.

The Second Dahlquist Barrier

I searched for a long time, finally Professor Lax showed me the Riesz-Herglotz theorem and I knew that I had my theorem.

(G. Dahlquist 1979)

Theorem 1.3. *If the multistep method (1.1) is A -stable, then*

$$\operatorname{Re} \left(\frac{\varrho(\zeta)}{\sigma(\zeta)} \right) > 0 \quad \text{for} \quad |\zeta| > 1. \quad (1.18)$$

For irreducible methods the converse is also true: (1.18) implies A -stability.

Proof. If the method is A -stable then all roots of (1.6) must satisfy $|\zeta| \leq 1$ whenever $\operatorname{Re} \mu \leq 0$. The logically equivalent statement ($\operatorname{Re} \mu > 0$ whenever $|\zeta| > 1$) yields (1.18) since by (1.6) $\mu = \varrho(\zeta)/\sigma(\zeta)$.

Suppose now that (1.18) holds and that the method is irreducible. Fix a μ_0 with $\operatorname{Re} \mu_0 \leq 0$ and let ζ_0 be a root of (1.6). We then have $\sigma(\zeta_0) \neq 0$ (otherwise the method would be reducible). Hence $\mu_0 = \varrho(\zeta_0)/\sigma(\zeta_0)$ and it follows from (1.18) that $|\zeta_0| \leq 1$. We still have to show that ζ_0 is a simple root if $|\zeta_0| = 1$. By a continuity argument it follows from (1.18) that $|\zeta_0| = 1$ and $\operatorname{Re} \mu_0 < 0$ are contradictory. Therefore, it remains to prove that for $\operatorname{Re} \mu_0 = 0$ a root satisfying $|\zeta_0| = 1$ must be simple. In a neighbourhood of such a root we have

$$\frac{\varrho(\zeta)}{\sigma(\zeta)} - \mu_0 = C_1(\zeta - \zeta_0) + C_2(\zeta - \zeta_0)^2 + \dots$$

and (1.18) implies that $C_1 \neq 0$. This, however, is only possible if ζ_0 is a simple root of (1.6). \square

In all the above examples we have not yet seen an A -stable multistep formula of order $p \geq 3$. The following famous theorem explains this observation.

Theorem 1.4 (Dahlquist 1963). *An A -stable multistep method must be of order $p \leq 2$. If the order is 2, then the error constant satisfies*

$$C \leq -\frac{1}{12}. \quad (1.19)$$

The trapezoidal rule is the only A -stable method of order 2 with $C = -1/12$.

Proof. Dahlquist's first proof of this theorem is difficult. More elementary versions emerged in Widlund (1967), in lecture notes of W. Liniger (Univ. of Neuchâtel 1971) and in the book of Grigorieff (1977, vol.2, p. 218).

We start by recalling some formulas from Volume I: Eq. (ii) of Theorem III.2.4 and Eq. (III.2.7) yield

$$\varrho(e^h) - h\sigma(e^h) = C_{p+1}h^{p+1} + \dots \quad \text{for } h \rightarrow 0. \quad (1.20)$$

From the consistency conditions (III.2.6) we have

$$\varrho(e^h) = \varrho(1 + h + \dots) = \varrho(1) + \varrho'(1)h + \dots = \sigma(1)h + \dots$$

We divide (1.20) by $h\varrho(e^h)$ and obtain

$$\frac{1}{h} - \frac{\sigma(e^h)}{\varrho(e^h)} = Ch^{p-1} + \dots \quad \text{for } h \rightarrow 0 \quad (1.21)$$

where C is the *error constant* (III.2.13). With $\zeta = e^h$ this becomes

$$\frac{1}{\log \zeta} - \frac{\sigma(\zeta)}{\varrho(\zeta)} = C(\zeta - 1)^{p-1} + \dots \quad \text{for } \zeta \rightarrow 1. \quad (1.22)$$

In this formula we put $p = 2$. Whenever the method is of higher order, we have $C = 0$. When the order of the method is one, we have nothing to prove. The same formula for the trapezoidal rule for which $\varrho_T(\zeta) = \zeta - 1$, $\sigma_T(\zeta) = \frac{1}{2}(\zeta + 1)$, becomes by series expansion (or by using Table III.2.1)

$$\frac{1}{\log \zeta} - \frac{\sigma_T(\zeta)}{\varrho_T(\zeta)} = -\frac{1}{12}(\zeta - 1) + \dots \quad \text{for } \zeta \rightarrow 1. \quad (1.23)$$

The idea is now to subtract the two formulas and obtain

$$d(\zeta) := \frac{\sigma(\zeta)}{\varrho(\zeta)} - \frac{\sigma_T(\zeta)}{\varrho_T(\zeta)} = \left(-C - \frac{1}{12}\right)(\zeta - 1) + \dots \quad \text{for } \zeta \rightarrow 1. \quad (1.24)$$

From (1.18) we have that

$$\operatorname{Re} \left(\frac{\varrho(\zeta)}{\sigma(\zeta)} \right) > 0 \quad \text{or equivalently} \quad \operatorname{Re} \left(\frac{\sigma(\zeta)}{\varrho(\zeta)} \right) > 0 \quad \text{for } |\zeta| > 1. \quad (1.25)$$

The point here is that for the trapezoidal rule this $\operatorname{Re}(\dots)$ is *zero* for $|\zeta| = 1$ since this method has precisely \mathbb{C}^- as stability domain. Hence from (1.24) we obtain

$$\lim_{\substack{\zeta \rightarrow \zeta_0 \\ |\zeta| > 1}} \operatorname{Re} d(\zeta) \geq 0 \quad \text{for } |\zeta_0| = 1. \quad (1.26)$$

The poles of $d(\zeta)$ are the roots of $\varrho(\zeta)$, which, by stability, are not allowed outside the unit circle. Thus, by the maximum principle, (1.26) remains true everywhere outside the unit circle. Choosing then $\zeta = 1 + \varepsilon$ with $\operatorname{Re} \varepsilon > 0$ and $|\varepsilon|$ small, we see from (1.24) that either $-C - \frac{1}{12} > 0$ or $d(\zeta) \equiv 0$. This concludes the proof. \square

Exercises

1. The Milne-Simpson methods for $k = 4$ and 5 satisfy $\operatorname{Re}(\varrho(\zeta)/\sigma(\zeta)) \geq 0$ for $|\zeta| = 1$. Since their order is higher than 2 , this seems to be in contradiction with the above proof of Theorem 1.4. Explain.
2. For the explicit midpoint rule (1.15), do the endpoints $\pm i$ of the stability region belong to S ? Study the (possible) stability of this method applied with $h = 1$ to $u' = v$, $v' = -u$.
3. Compute for the explicit and implicit Adams methods the largest $\lambda_0 \in \mathbb{R}$ such that the real interval $[-\lambda_0, 0]$ lies in S . Show that for the k -step explicit Adams methods we have $\lambda_0 = 2/u_k$ with $u_k = \sum_{j=0}^{k-1} 2^j \gamma_j$ ($u_1 = 1$, $u_2 = 2$, $u_3 = 11/3$, $u_4 = 20/3$, $u_5 = 551/45, \dots$). The use of generating functions (see Sect. III.1) allow us to show that

$$\sum_{j=1}^{\infty} u_k t^k = \left(-1 + \frac{2}{1-t} - \frac{1}{1-2t} \right) \log(1-2t),$$

a series with convergence radius $1/2$. This explains why these stability domains decrease so rapidly.

Hint. Just set $\theta = \pi$ in the root locus curve.

4. Prove that the stability region of the k -step, implicit Adams methods is of finite size for every $k \geq 2$.

Hint. Show that $(-1)^k \sigma(-1) < 0$, so that σ has a real negative root, smaller than -1 .

5. a) Show that all 2-step methods of order 2 are given by

$$\begin{aligned}\varrho(\zeta) &= (\zeta - 1)(\alpha\zeta + 1 - \alpha) \\ \sigma(\zeta) &= (\zeta - 1)^2\beta + (\zeta - 1)\alpha + (\zeta + 1)/2\end{aligned}$$

(which are irreducible for $\alpha \neq 2\beta$).

b) The method is stable at 0 iff $\alpha \geq 1/2$.

c) The method is stable at ∞ iff

$$\alpha \geq 1/2 \quad \text{and} \quad \beta > \alpha/2. \quad (1.27)$$

Apply the Schur-Cohn criterion (Sect. III.3, Exercise 4).

d) The method is A -stable iff (1.27) holds.

Hint.

$$\frac{\sigma(\zeta)}{\varrho(\zeta)} = \frac{1}{2} \cdot \frac{\zeta + 1}{\zeta - 1} + \left(\beta - \frac{\alpha}{2} \right) \cdot \frac{\zeta - 1}{\alpha\zeta + 1 - \alpha}.$$

V.2 “Nearly” A-Stable Multistep Methods

We are not attempting to disprove Dahlquist’s theorems but are trying to get round the conditions they impose . . .

(J. Cash 1979)

Dahlquist’s condition $p \leq 2$ for the order of an A -stable linear multistep method is a severe restriction for efficient practical calculations of high precision. There are only two ways of “breaking” this barrier:

- either weaken the condition;
- or strengthen the method.

These two points will occupy our attention in this and in the following section.

$A(\alpha)$ -Stability and Stiff Stability

It is the purpose of this note to show that a slightly different stability requirement permits methods of higher accuracy.

(O. Widlund 1967)

The angle α is only one of a number of parameters which have been proposed for measuring the extent of the stability region. But it is probably the best such measure . . .

(Skeel & Kong 1977)

Many important classes of practical problems do not require stability on the entire left half-plane \mathbb{C}^- . Further, for eigenvalues on the imaginary axis, the solutions are often highly oscillatory and one is then forced anyhow to restrict the step size “to the highest frequency present in order to represent the signal” (Gear 1971, p. 214).

Definition 2.1 (Widlund 1967). A convergent linear multistep method is $A(\alpha)$ -stable, $0 < \alpha < \pi/2$, if

$$S \supset S_\alpha = \{\mu ; |\arg(-\mu)| < \alpha, \mu \neq 0\}. \quad (2.1)$$

A method is $A(0)$ -stable if it is $A(\alpha)$ -stable for some (sufficiently small) $\alpha > 0$.

Similarly, Gear (1971) required in his famous concept of “*stiff stability*” that

$$S \supset \{\mu ; \operatorname{Re} \mu < -D\} \quad (2.2)$$

for some $D > 0$ and that the method be “accurate” in a rectangle $-D \leq \operatorname{Re} \mu \leq a$, $-\theta \leq \operatorname{Im} \mu \leq \theta$ for some $a > 0$ and θ about $\pi/5$. Many subsequent writers

didn’t like the inaccurate meaning of “accurate” in this definition and replaced it by something else. For example Jeltsch (1976) required that in addition to (2.2),

$$|\zeta_1(\mu)| > |\zeta_i(\mu)|, \quad i = 2, \dots, k \quad \text{in} \quad |\operatorname{Re} \mu| \leq a, \quad |\operatorname{Im} \mu| \leq \theta, \quad (2.3)$$

where $\zeta_1(\mu)$ is the analytic continuation of the principal root $\zeta_1(0) = 1$ of (1.6). Also, the rectangle given by

$$|\operatorname{Im} \mu| \leq \theta, \quad -D \leq \operatorname{Re} \mu \leq -a$$

should belong to S .

Other concepts are A_0 -stable (Cryer 1973) if

$$|\zeta_i(x)| < 1, \quad i = 1, \dots, k \quad \text{for} \quad -\infty < x < 0 \quad (2.4)$$

and \check{A} -stable (a joke of O. Nevanlinna 1979) if

$$(-\infty, 0] \subset S. \quad (2.5)$$

Of course, we have

$$A(0)\text{-stable} \implies A_0\text{-stable} \implies \check{A}\text{-stable} \quad (2.6)$$

but neither implication is reversible (Exercise 3; see also “Theorem 1” of Jeltsch 1976).

The BDF methods (1.18) satisfy (2.1) for $A(\alpha)$ -stability and (2.2) for stiff stability with the values

k	1	2	3	4	5	6	
α	90°	90°	86.03°	73.35°	51.84°	17.84°	(2.7)
D	0	0	0.083	0.667	2.327	6.075	

High Order $A(\alpha)$ -Stable Methods

Dill and Gear . . . and Jain and Srivastava . . . have used computers to construct stiffly stable methods of orders eight and eleven, respectively, but were unable to construct higher order stiffly stable methods. Even though we have shown here that A_0 -stable methods of arbitrarily high order exist, we conjecture that $A(0)$ -stable linear multistep methods of higher order, of order greater than 20 say, do not exist. (Cryer 1973)

Widlund (1967) showed that for every $\alpha < \pi/2$, α arbitrarily close to $\pi/2$, there exist $A(\alpha)$ -stable multistep methods of order $p = k$ for $p = 3$ and $p = 4$. It is now an interesting question whether such methods also exist for higher orders. Well, the answer consists of good news and bad news.

First the good news. The conjecture of Cryer (see quotation) was quickly disproved by combining Cryer’s A_0 -stable methods with the result of Jeltsch (1976)

which says that certain A_0 -stable methods are also $A(\alpha)$ -stable. The following theorem shows that α can even be chosen arbitrarily close to $\pi/2$:

Theorem 2.2 (Grigorieff & Schroll 1978). *Let $\alpha < \pi/2$ be given. Then for every $k \in \mathbb{N}$ there exists an $A(\alpha)$ -stable linear k -step method of order $p = k$.*

Proof. For $p = k = 2$ the two-step BDF method which is A -stable, and hence $A(\alpha_2)$ -stable for every $\alpha_2 \leq \pi/2$, does the job. For k arbitrary, we intercalate $k - 2$ values between α and $\pi/2$,

$$\alpha < \alpha_{k-1} < \alpha_{k-2} < \dots < \alpha_3 < \alpha_2 \leq \frac{\pi}{2}, \quad (2.8)$$

and extend the method step by step with the help of Lemma 2.3. \square

Lemma 2.3. *Suppose an $A(\alpha)$ -stable k -step method of order p is given with*

$$\varrho(\zeta) \neq 0 \quad \text{if } |\zeta| = 1, \zeta \neq 1 \quad (2.9a)$$

$$\sigma(\zeta) \neq 0 \quad \text{if } |\zeta| = 1. \quad (2.9b)$$

Then for every $\tilde{\alpha} < \alpha$ there exists an $A(\tilde{\alpha})$ -stable $(k+1)$ -step method of order $p+1$ which also satisfies (2.9).

The *proof* follows very closely the ideas of Jeltsch & Nevanlinna (1982): Let $\varrho(\zeta)$ and $\sigma(\zeta)$ represent the given k -step method with order condition

$$\frac{\varrho(\zeta)}{\log \zeta} - \sigma(\zeta) = C_{p+1}(\zeta - 1)^p + \mathcal{O}((\zeta - 1)^{p+1}). \quad (2.10)$$

If we multiply ϱ and σ by $(\zeta - 1)$ we formally increase the order by 1 and at the same time leave the root locus curve unchanged. Everything seems to be proved. However, the new ϱ -polynomial would have a double root at $\zeta = 1$ and would thus lead to an unstable method. We therefore choose $\varepsilon > 0$ and multiply (2.10) by $(\zeta - 1 + \varepsilon)$, which moves the root slightly inside the unit circle. We then obtain a new method of order $p+1$ if we put

$$\begin{aligned} \tilde{\varrho}(\zeta) &= \varrho(\zeta)(\zeta - 1 + \varepsilon) \\ \tilde{\sigma}(\zeta) &= \sigma(\zeta)(\zeta - 1 + \varepsilon) + \varepsilon C_{p+1}(\zeta - 1)^p. \end{aligned} \quad (2.11)$$

Since $p = k + 2$ is excluded (by Theorem III.3.9 methods with $p = k + 2$ are symmetric and violate Hypothesis (2.9a)), both polynomials $\tilde{\varrho}$ and $\tilde{\sigma}$ are of degree $\leq k + 1$. Now the formula

$$\frac{\tilde{\sigma}(\zeta)}{\tilde{\varrho}(\zeta)} - \frac{\sigma(\zeta)}{\varrho(\zeta)} = \frac{\varepsilon C_{p+1}(\zeta - 1)^p}{\varrho(\zeta)(\zeta - 1 + \varepsilon)} \quad (2.12)$$

allows us to compare, for ε small, the root-locus curves of the two methods. The fact that we are working with $\sigma(e^{i\theta})/\varrho(e^{i\theta}) = 1/\mu$ instead of $\mu = \varrho(e^{i\theta})/\sigma(e^{i\theta})$

does not matter, because the transformation $\mu \mapsto 1/\mu$ maps the sector of Definition 2.1 onto itself. Because of Hypothesis (2.9a), 1 is the only (simple) root of $\varrho(\zeta)$ on the unit circle, therefore

$$\left| \frac{\tilde{\sigma}(\zeta)}{\tilde{\varrho}(\zeta)} - \frac{\sigma(\zeta)}{\varrho(\zeta)} \right| \leq C \cdot \varepsilon \frac{|\zeta - 1|^{p-1}}{|\zeta - 1 + \varepsilon|} \quad \text{for } \zeta = e^{i\theta}. \quad (2.13)$$

A small obstacle still separates us from “endless pleasure, endless love, Semele enjoys above”: the denominator $|\zeta - 1 + \varepsilon|$, which becomes small for $\varepsilon \rightarrow 0$ and $\theta \rightarrow 0$. For $p > 1$, this “small” denominator is simply balanced by one of the factors $|\zeta - 1|$ from the numerator and we have

$$\left| \frac{\tilde{\sigma}(\zeta)}{\tilde{\varrho}(\zeta)} - \frac{\sigma(\zeta)}{\varrho(\zeta)} \right| \leq \hat{C} \cdot \varepsilon \quad (2.14)$$

which means uniform pointwise convergence of $\tilde{\sigma}(\zeta)/\tilde{\varrho}(\zeta)$ to $\sigma(\zeta)/\varrho(\zeta)$ if $\varepsilon \rightarrow 0$. Since $\sigma(\zeta)/\varrho(\zeta)$ is bounded away from the origin (Hypothesis (2.9b)), this also means uniform convergence of the angles.

This is already sufficient to prove Theorem 2.2, where we always have $p \geq 2$. However, Lemma 2.3 remains valid for $p = 1$ too: the critical region is when $\theta \rightarrow 0$, in which case $|\sigma(e^{i\theta})/\varrho(e^{i\theta})|$ and $|\tilde{\sigma}(e^{i\theta})/\tilde{\varrho}(e^{i\theta})|$ tend to infinity like Const/θ . Instead of (2.14) we have for $p = 1$

$$\left| \frac{\tilde{\sigma}(\zeta)}{\tilde{\varrho}(\zeta)} - \frac{\sigma(\zeta)}{\varrho(\zeta)} \right| \leq \frac{C\varepsilon}{|\zeta - 1 + \varepsilon|} = \mathcal{O}\left(\frac{\varepsilon}{\theta}\right).$$

Thus the *angle* (seen from the origin) between $\tilde{\sigma}(\zeta)/\tilde{\varrho}(\zeta)$ and $\sigma(\zeta)/\varrho(\zeta)$ is $\mathcal{O}(\varepsilon)$. \square

Approximating Low Order Methods with High Order Ones

The above proof of Lemma 2.3 actually shows more than angle-boundedness of the root locus curve, namely uniform convergence of the root locus curve of a high order method to that of a lower order one. This leads to the following theorem of Jeltsch & Nevanlinna (1982):

Theorem 2.4. *Let a linear stable k -step method of order p and stability domain S be given which satisfies (2.9a). Then to any closed set $\Omega \subset \text{Int } S \subset \overline{\mathbb{C}}$ and any $K \in \mathbb{N}$ there exists a linear $k + K$ -step method of order $p + K$ whose stability domain \tilde{S} satisfies*

$$\tilde{S} \supset \Omega.$$

Moreover if the first method is explicit, the higher-order method is also explicit.

Proof. The proof is similar to that of Lemma 2.3. Instead of the sequence (2.8) we use a sequence of embedded closed and open subsets between Ω and S (Urysohn’s Lemma). Hypothesis (2.9b) is ruled out by passing to the compactified topology of $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. \square

Remark. No method with non-empty $\text{Int } S$ of practical interest violates Hypothesis (2.9a). Nevertheless, Theorem 2.4 remains valid *without* this hypothesis, but the proof becomes more complicated (see “Lemma 3.6” of Jeltsch & Nevanlinna 1982).

A Disc Theorem

Another weakening of A -stability is to require stability for

$$D_r = \{\mu ; |\mu + r| \leq r\}, \quad (2.15)$$

which is a disc of radius r in \mathbb{C}^- tangent to the imaginary axis at the origin. Theorems about stability in D_r are stronger than theorems about $A(\alpha)$ -stability for eigenvalues close to the origin. The following result is, again, due to Jeltsch & Nevanlinna (1982):

Theorem 2.5. *Let a linear k -step method of order p be given with $S \supset D_r$. Then for any $\tilde{r} < r$ and any $K \in \mathbb{N}$ there exists a linear $k + K$ -step method of order $p + K$ whose stability domain \tilde{S} satisfies $\tilde{S} \supset D_{\tilde{r}}$.*

Proof. The map $\mu \mapsto 1/\mu$ used in the proof of Lemma 2.3 maps the exterior of D_r onto the half-plane

$$\left\{ \mu \in \mathbb{C} ; \text{Re } \mu > -\frac{1}{2r} \right\}. \quad (2.16)$$

Therefore the uniform convergence established in (2.14) also covers the new situation if $p > 1$. The case $p = 1$, however, needs a more careful study and we refer to the original paper of Jeltsch & Nevanlinna (1982, pp. 277-279). \square

Accuracy Barriers for Linear Multistep Methods

Now here is the “bad news”: high order $A(\alpha)$ -stable methods, for α close to $\pi/2$, cannot be of practical use, or in other words: “the second Dahlquist barrier cannot be broken”. The reason is simply that high order alone is not sufficient for high accuracy, because the methods then have enormous error constants. Jeltsch & Nevanlinna (1982) give an impressive staccato (from “Theorem 4.1” to “Lemma 4.15”) of lower bounds for error constants and Peano kernels of methods having large stability domains. The Peano kernels, the most serious measures for the error, are defined by the formulas (see (III.2.15) and (III.2.3) of Volume I)

$$L(x) = h^{q+1} \int_{-\infty}^{\infty} \tilde{K}_q(-s) y^{(q+1)}(x + sh) ds \quad (2.17)$$

$$= \sum_{j=0}^k (\alpha_j y(x + jh) - h\beta_j y'(x + jh)). \quad (2.18)$$

The kernels $\tilde{K}_q(-s) = K_q(s)$ are zero outside the interval $0 \leq s \leq k$ and are piecewise polynomials given by complicated formulas (see (III.2.16)) which appear not very attractive to work with.

However, the formulas simplify if we use the *Fourier transform* which, for a function $f(x)$, is defined by

$$\widehat{f}(\xi) = \int_{-\infty}^{\infty} e^{-ix\xi} f(x) dx. \quad (2.19)$$

We obtain \widehat{L} from (2.17) by insertion of the definitions, several integrations by parts and transformations of double integrals:

$$\widehat{L}(\xi) = h^{q+1} \widehat{\tilde{K}_q}(h\xi) \cdot \widehat{y^{(q+1)}}(\xi) \quad (2.20)$$

$$= \widehat{\tilde{K}_q}(h\xi) (ih\xi)^{q+1} \widehat{y}(\xi), \quad (2.21)$$

and from (2.18)

$$\widehat{L}(\xi) = (\varrho(e^{ih\xi}) - ih\xi \sigma(e^{ih\xi})) \cdot \widehat{y}(\xi). \quad (2.22)$$

Thus (2.20) and (2.22) give

$$\widehat{\tilde{K}_q}(-\xi) = \widehat{\tilde{K}_q}(\xi) = (\varrho(e^{i\xi}) - i\xi \sigma(e^{i\xi}))(i\xi)^{-(q+1)}, \quad (2.23)$$

a nice formula, involving the polynomials ϱ and σ , with which we are better acquainted.

What about the usefulness of $\widehat{\tilde{K}_q}$ for error estimates? Well, it is the *Parseval identity* (Exercise 4)

$$\|f\|_{L^2(-\infty, \infty)} = \frac{1}{\sqrt{2\pi}} \|\widehat{f}\|_{L^2(-\infty, \infty)} \quad (2.24)$$

which allows us to obtain the L^2 -estimate for the error

$$\|L\|_{L^2(-\infty, \infty)} \leq h^{q+1} \|\widehat{\tilde{K}_q}\|_{L^\infty} \cdot \|y^{(q+1)}\|_{L^2}, \quad (2.25)$$

as follows:

$$\begin{aligned} \|L\|_{L^2(-\infty, \infty)}^2 &= \frac{1}{2\pi} \|\widehat{L}\|_{L^2(-\infty, \infty)}^2 && \text{(from (2.24))} \\ &= \frac{h^{2q+2}}{2\pi} \int_{-\infty}^{\infty} |\widehat{\tilde{K}_q}(\xi)|^2 |\widehat{y^{(q+1)}}(\xi)|^2 d\xi && \text{(from (2.20))} \\ &\leq \frac{h^{2q+2}}{2\pi} \max |\widehat{\tilde{K}_q}(\xi)|^2 \cdot \int_{-\infty}^{\infty} |\widehat{y^{(q+1)}}(\xi)|^2 d\xi && \text{(estimation)} \\ &= \frac{h^{2q+2}}{2\pi} \|\widehat{\tilde{K}_q}\|_{L^\infty}^2 \cdot \|y^{(q+1)}\|_{L^2}^2 && \text{(definitions)} \\ &= h^{2q+2} \|\widehat{\tilde{K}_q}\|_{L^\infty}^2 \cdot \|y^{(q+1)}\|_{L^2}^2. && \text{(from (2.23), (2.24))} \end{aligned}$$

In order that the obtained estimates (2.25) for L express the *actual errors* of the numerical solution, we adopt throughout this section the normalization $\sigma(1) = 1$ (cf. Eq. (III.2.13)).

And here is the theorem which tells us that linear multistep methods of order $p > 2$ and “large” stability domain cannot be precise:

Theorem 2.6 (Jeltsch & Nevanlinna 1982). *Consider k -step methods of order $p > 2$, normalized by $\sigma(1) = 1$, for which the disc D_r of (2.15) is in the stability domain S . Then there exists a constant $C > 0$ (depending on k, p, q ; but independent of r) such that the Fourier transform of the Peano kernel K_q ($q \leq p$) satisfies*

$$\|\widehat{K}_q\|_\infty \geq C \left(\frac{r}{3}\right)^{p-2}. \quad (2.26)$$

The proof of Jeltsch & Nevanlinna is in two steps:

- a) The stability requirement forces some coefficients a_j of $R(z)$ to be large (Lemma 2.7 below), where as in (III.3.17)

$$R(z) = \left(\frac{z-1}{2}\right)^k \varrho\left(\frac{z+1}{z-1}\right) = \sum_{j=0}^k a_j z^j \quad (2.27)$$

$$S(z) = \left(\frac{z-1}{2}\right)^k \sigma\left(\frac{z+1}{z-1}\right) = \sum_{j=0}^k b_j z^j. \quad (2.28)$$

- b) $\|\widehat{K}_q\|_{L^\infty}$ can be bounded from below by $\max_j a_j$ (Lemma 2.8).

Lemma 2.7. *If $D_r \subset S$ and $p > 2$ then*

$$a_{k-j} \geq \left(\frac{r}{3}\right)^{j-1} \cdot a_{k-1} = \left(\frac{r}{3}\right)^{j-1} \cdot 2^{1-k} \quad \text{for } j = 2, \dots, p-1. \quad (2.29)$$

Proof. Stability in D_r means that for $\mu \in D_r$ all roots of $\varrho(\zeta) - \mu\sigma(\zeta) = 0$ lie in $|\zeta| \leq 1$. Hence

$$\varrho(\zeta)/\sigma(\zeta) \notin D_r \quad \text{for } |\zeta| > 1. \quad (2.30)$$

Applying the Graeco-Roman transformation $\zeta = (z+1)/(z-1)$ and using (2.16) this means that

$$\operatorname{Re} \frac{S(z)}{R(z)} > -\frac{1}{2r} \quad \text{for } \operatorname{Re} z > 0 \quad (2.31)$$

or

$$\operatorname{Re} \frac{2rS(z) + R(z)}{R(z)} > 0 \quad \text{for } \operatorname{Re} z > 0. \quad (2.32)$$

Next, we must consider the order conditions (Lemma III.3.7 and Exercise 9 of Sect. III.3)

$$R(z) \left(\frac{z}{2} - \frac{1}{6z} - \frac{2}{45z^3} - \dots \right) - S(z) = \mathcal{O}\left(\left(\frac{1}{z}\right)^{p-k}\right), \quad z \rightarrow \infty. \quad (2.33)$$

This shows that $R(z) = \mathcal{O}(z^{k-1})$, $S(z) = \mathcal{O}(z^k)$, but $2S(z) - zR(z) = \mathcal{O}(z^{k-1})$. Thus we subtract rz from (2.32) in order to lower the degree of the numerator. The

resulting function again satisfies

$$\operatorname{Re} \frac{r(2S(z) - zR(z)) + R(z)}{R(z)} > 0 \quad \text{for } \operatorname{Re} z > 0 \quad (2.34)$$

because of $\operatorname{Re}(rz) = 0$ on $z = iy$ and the maximum principle (an idea similar to that of Lemma IV.5.21). The function (2.34) can therefore have no zeros in \mathbb{C}^+ (since by Taylor expansion all arguments of a function appear in a complex neighbourhood of a zero). Therefore the *numerator* of (2.34) must have non-negative coefficients (cf. the proof of Lemma III.3.6). Multiplying out (2.33) and (2.34) we obtain for the coefficient of z^{k-j} ($j \leq p-1$):

$$0 \leq r \left(-\frac{1}{3} a_{k-j+1} - \frac{4}{45} a_{k-j+3} - \dots \right) + a_{k-j}$$

or by simplifying (cf. Lemmas III.3.8 and III.3.6)

$$\frac{r}{3} a_{k-j+1} \leq a_{k-j}.$$

Using $a_{k-1} = 2^{1-k} \varrho'(1) = 2^{1-k}$ (see Lemma III.3.6), this leads to (2.29). \square

Lemma 2.8. *There exists $C > 0$ (depending on k, p and q with $q = 0, 1, \dots, p$) with the following property: if $0 \in S$, then*

$$\|\widehat{K}_q\|_{L^\infty} \geq C \cdot \max_j a_j. \quad (2.35)$$

Proof. We set $e^{i\xi} = \zeta$, $\xi = -i \log \zeta$ in Eq. (2.23) so that the maximum must be taken over the set $|\zeta| = 1$. Then we introduce $\zeta = (z+1)/(z-1)$ and take the maximum over the imaginary axis. This gives with (2.27) and (2.28)

$$\|\widehat{K}_q\|_{L^\infty} = \sup_t \underbrace{\left| \frac{1}{(it)^k} \left(\frac{R(it)}{\log \frac{it+1}{it-1}} - S(it) \right) \right|}_{\Phi(t)} \cdot \underbrace{\left| \left(\frac{2it}{it-1} \right)^k \cdot \left| \log \left(\frac{it+1}{it-1} \right) \right|^{-q} \right|}_{\Psi(t)}. \quad (2.36)$$

We now insert, for $|t| > 1$, Eqs. (III.3.19), (III.3.21) and (III.3.22) to obtain

$$|\Phi(t)| = \left| P_k \left(\frac{1}{it} \right) + \frac{d_1}{(it)^{k+1}} + \frac{d_2}{(it)^{k+2}} + \dots \right| \quad (2.37)$$

where P_k is a polynomial of degree k and subdegree p (see Lemma III.3.7), determined by the method. Since we want our estimates to be true for *all* methods, we treat P_k as an *arbitrary* polynomial. Separating real and imaginary parts and substituting $1/t = s$ gives

$$\begin{aligned} |\Phi(t)|^2 &= |Q_{k-1}(s) + d_1 s^{k+1} - d_3 s^{k+3} + \dots|^2 \\ &\quad + |Q_k(s) + d_2 s^{k+2} - d_4 s^{k+4} + \dots|^2 = |\Phi_1(t)|^2 + |\Phi_2(t)|^2 \end{aligned} \quad (2.38)$$

where $Q_{k-1}(s)$ and $Q_k(s)$ are arbitrary (even or odd) polynomials of subdegree p and degree $k-1$ and k , respectively. Both terms are minorized separately, e.g. for the first we write

$$|\Phi_1(t)| \geq |Q_{k-1}(s) + d_1 s^{k+1}| - |d_3 s^{k+3} - d_5 s^{k+5} + \dots|. \quad (2.39)$$

Since $\mu_1 < \mu_3 < \mu_5 < \dots < 0$ (Exercise 6 below) and $a_i \geq 0$ we have from (III.3.22)

$$d_1 \leq d_3 \leq d_5 \leq \dots \leq 0 \quad \text{and} \quad d_2 \leq d_4 \leq d_6 \leq \dots \leq 0. \quad (2.40)$$

Therefore, the second term in (2.39) is majorized by the alternating series argument for $0 < s < 1$ as

$$|d_3 s^{k+3} - d_5 s^{k+5} + \dots| \leq |d_3| s^{k+3} \leq |d_1| s^{k+3}.$$

Since $Q_{k-1}(s)$ is an arbitrary polynomial, we can replace it by $|d_1|Q_{k-1}(s)$ so that $|d_1|$ becomes a common factor of the whole expression

$$|\Phi_1(t)| \geq |d_1| \left(|Q_{k-1}(s) + s^{k+1}| - s^{k+3} \right). \quad (2.41)$$

This suggests that we define the constants

$$\begin{aligned} D_1 &= \inf_{Q_{k-1}} \left\{ \sup_{0 < s < 1} \left[\left(|Q_{k-1}(s) + s^{k+1}| - s^{k+3} \right) \left(\frac{2}{\sqrt{1+s^2}} \right)^k \left(\frac{1}{2 \arctan s} \right)^q \right] \right\} \\ D_2 &= \inf_{Q_k} \left\{ \sup_{0 < s < 1} \left[\left(|Q_k(s) + s^{k+2}| - s^{k+4} \right) \left(\frac{2}{\sqrt{1+s^2}} \right)^k \left(\frac{1}{2 \arctan s} \right)^q \right] \right\} \end{aligned} \quad (2.42)$$

where the \inf is taken over all polynomials $Q_{k-1}(s) = c_{k-1}s^{k-1} + c_{k-3}s^{k-3} + c_{k-5}s^{k-5} + \dots$ respectively $Q_k(s) = c_k s^k + c_{k-2}s^{k-2} + c_{k-4}s^{k-4} + \dots$ of subdegree p . The last two factors represent $\Psi(t)$ of (2.36). Since s^{k+1} dominates s^{k+3} for small s , D_1 and D_2 are *positive* constants (see Exercise 8). We then have from (2.38) and (2.36)

$$\|\widehat{K}_q\|_{L^\infty} \geq \sqrt{d_1^2 D_1^2 + d_2^2 D_2^2} \quad (2.43)$$

Since both d_1 and d_2 are sums of a_j with negative coefficients (see (III.3.22) and Lemma III.3.8), $\|\widehat{K}_q\|_\infty$ must be large if one of the coefficient a_j is large. \square

This concludes the proof of Theorem 2.6 which, by the way, also proves Theorem 1.4 again. \square

Exercises

1. Show that no explicit method can be $A(0)$ -stable.
2. Show that $\beta_k/\alpha_k > 0$ is a necessary condition for an $A(\alpha)$ -stable linear k -step method.
3. a) Show that the method

$$y_{n+2} - y_{n+1} = \frac{h}{4}(f_{n+2} + 2f_{n+1} + f_n)$$

has a stability domain bounded by a parabola. It is therefore A_0 -stable, but not $A(0)$ -stable (Cryer 1973).

b) Find a “deformation” of the 5th order BDF scheme

$$\sum_{j=1}^5 \frac{1}{j} \nabla^j y_{n+1} + \beta \nabla^6 y_{n+1} = h f_{n+1}$$

with $\beta \approx 0.232 \dots$ which is \dot{A} -stable, but not A_0 -stable.

c) Find a method which is A_0 -stable, but not stable at infinity.

Hint for (c). If you “lift up your heads, o ye gates” (just a few lines, not to heaven), the answer is easy to find.

4. (Parseval 1799). Prove the identity (2.24).

Hint. Insert the definitions into

$$\|\hat{f}\|_{L^2}^2 = \int_{-\infty}^{\infty} \hat{f}(\xi) \overline{\hat{f}(\xi)} d\xi$$

to get a triple integral. Two of these integrals then disappear with the Fourier inversion formula.

Remark. You may be astonished to see that Parseval’s identity is older than Fourier series and Fourier transforms. Well, Parseval’s identity was originally a formula between an infinite sum and an integral, which was later re-interpreted and generalized to become what it is today.

5. Substitute $\xi = \pi$ in Formula (2.23) to obtain an easy minorization for $\|\widehat{K}_q\|_{L^\infty}$. Then compute for the methods defined in the proof of Lemma 2.3 (normalized by $\sigma(1) = 1$) the value $\sigma(-1)$ for ε small. This then shows that \widehat{K}_q becomes very large.
6. Use the formula (see the proof of Lemma III.3.8)

$$\mu_{2j+1} = \int_{-1}^{+1} x^{2j} \left(\left(\log \frac{1+x}{1-x} \right)^2 + \pi^2 \right)^{-1} dx$$

to show that $\mu_1 > \mu_3 > \mu_5 > \dots > 0$.

7. Show that for $q = p$ Eq. (2.23) becomes, by substituting $i\xi = h$ and letting $h \rightarrow 0$ in Eq. (1.20), $\widehat{K}_p(0) = C_{p+1}$, where C_{p+1} is, for $\sigma(1) = 1$, the *error constant*.

Formula (2.36) then provides, for $p = k$ and $t \rightarrow \infty$, lower bounds for the error constant (see “Theorem 4.5” of Jeltsch & Nevanlinna 1982).

8. For $p = k + 1$, the polynomials Q_{k-1} and Q_k in (2.42) vanish identically, because the subdegree must be p . Compute in this case the constants D_1 and D_2 . It is also easy to compute them for $p = k - 1$. In the general case the optimal solution satisfies a sort of “Tchebysheff alternative”.

Results.

Case $p = k + 1$ ($Q = 0$):

D_1	$p=3$ $k=2$	$p=4$ $k=3$	$p=5$ $k=4$	$p=6$ $k=5$	D_2	$p=3$ $k=2$	$p=4$ $k=3$	$p=5$ $k=4$	$p=6$ $k=5$
$q=0$	0.4742	0.5695	0.7020	0.8813	$q=0$	0.3607	0.4501	0.5706	0.7319
$q=1$	0.3876	0.4435	0.5298	0.6505	$q=1$	0.2754	0.3347	0.4163	0.5263
$q=2$	0.3524	0.3659	0.4152	0.4933	$q=2$	0.2205	0.2570	0.3108	0.3852
$q=3$	0.5000	0.3381	0.3459	0.3891	$q=3$	0.1935	0.2075	0.2400	0.2888
$q=4$		0.5000	0.3251	0.3275	$q=4$		0.1849	0.1956	0.2244
$q=5$			0.5000	0.3131	$q=5$			0.1770	0.1845
$q=6$				0.5000	$q=6$				0.1698

Case $p = k - 1$ (one free constant in Q):

D_1	$p=3$ $k=4$	$p=4$ $k=5$	$p=5$ $k=6$	$p=6$ $k=7$	D_2	$p=3$ $k=4$	$p=4$ $k=5$	$p=5$ $k=6$	$p=6$ $k=7$
$q=0$	0.0511	0.0362	0.0262	0.0193	$q=0$	0.0195	0.0142	0.0104	0.0077
$q=1$	0.0727	0.0499	0.0353	0.0256	$q=1$	0.0269	0.0191	0.0138	0.0101
$q=2$	0.1100	0.0709	0.0486	0.0344	$q=2$	0.0384	0.0263	0.0186	0.0135
$q=3$	0.2031	0.1070	0.0691	0.0474	$q=3$	0.0583	0.0374	0.0256	0.0181
$q=4$		0.1962	0.1041	0.0673	$q=4$		0.0567	0.0365	0.0250
$q=5$			0.1894	0.1012	$q=5$			0.0552	0.0356
$q=6$				0.1828	$q=6$				0.0537

Case $p = k - 3$ (two free constants in Q):

D_1	$p=3$ $k=6$	$p=4$ $k=7$	$p=5$ $k=8$	$p=6$ $k=9$	D_2	$p=3$ $k=6$	$p=4$ $k=7$	$p=5$ $k=8$	$p=6$ $k=9$
$q=0$	0.0030	0.0014	0.0007	0.0003	$q=0$	0.0007	0.0004	0.0002	0.0001
$q=1$	0.0066	0.0029	0.0014	0.0007	$q=1$	0.0015	0.0007	0.0003	0.0002
$q=2$	0.0160	0.0066	0.0029	0.0014	$q=2$	0.0034	0.0015	0.0007	0.0003
$q=3$	0.0457	0.0158	0.0065	0.0029	$q=3$	0.0082	0.0034	0.0015	0.0007
$q=4$		0.0448	0.0156	0.0064	$q=4$		0.0081	0.0033	0.0015
$q=5$			0.0439	0.0154	$q=5$			0.0080	0.0033
$q=6$				0.0431	$q=6$				0.0079

V.3 Generalized Multistep Methods

The Dahlquist bound of two on the order of A -stable multistep methods was the imperative to propound . . . weaker stability properties, . . . An alternative approach for circumventing Dahlquist's bound is to modify the class of methods, rather than the property.

(T.A. Bickart & W.B. Rubin 1974)

The search for higher order A -stable multistep methods is carried out in two main directions:

- Use higher derivatives of the solutions;
- Throw in additional stages, off-step points, super-future points and the like, which leads into the large field of general linear methods.

Second Derivative Multistep Methods of Enright

Hermite's formulas are rediscovered and republished every four years.
(P.J. Davis 1963)

Differentiation of a differential equation

$$y' = f(x, y) \quad (3.1)$$

with respect to x gives the second derivative of the solution

$$y'' = f_x + f_y \cdot f =: g(x, y), \quad (3.2)$$

which we shall denote by g . Now a straightforward generalization of both multistep formulas (1.1) and, say, the Taylor series method (see I.8.13)

$$y_{n+1} = y_n + hf_n + \frac{h^2}{2!} g_n$$

can be written in the form

$$\sum_{i=0}^k \alpha_i y_{n+i} = h \sum_{i=0}^k \beta_i f_{n+i} + h^2 \sum_{i=0}^k \gamma_i g_{n+i} \quad (3.3)$$

where the α_i , β_i , γ_i are parameters which must be chosen appropriately. Most of the theory of linear multistep methods (Sect. III.2) generalizes without difficulty. Taylor expansion similar to (III.2.5) shows that method (3.3) is of order p if and only if

$$\sum_{i=0}^k \alpha_i i^q = q \sum_{i=0}^k \beta_i i^{q-1} + q(q-1) \sum_{i=0}^k \gamma_i i^{q-2} \quad (3.4)$$

for $0 \leq q \leq p$. The first two of these formulas are identical to (III.2.6), i.e., to

$$\varrho(1) = 0, \quad \varrho'(1) = \sigma(1). \quad (3.5)$$

The *error constant* (see Eq. (III.2.13) and Exercise 2 of Sect. III.4) is given by

$$C = \frac{1}{\sigma(1)(p+1)!} \left(\sum_{i=0}^k \alpha_i i^{p+1} - (p+1) \sum_{i=0}^k \beta_i i^p - (p+1)p \sum_{i=0}^k \gamma_i i^{p-1} \right). \quad (3.6)$$

A search for a good choice of the free parameters α_i , β_i , γ_i was undertaken by Enright (1974) with the following ideas:

- (i) Set $\alpha_k = 1$, $\alpha_{k-1} = -1$, $\alpha_{k-2} = \dots = \alpha_0 = 0$ to ensure reasonable stability in a neighbourhood of the origin as in the standard Adams formulas;
- (ii) Set $\gamma_k \neq 0$, $\gamma_{k-1} = \dots = \gamma_0 = 0$ to ensure stability at infinity as in the BDF formulas;
- (iii) Determine the remaining $k+2$ coefficients $\gamma_k, \beta_k, \beta_{k-1}, \dots, \beta_0$ from Equations (3.4) for $q = 1, 2, \dots, k+2$ ($q = 0$ is satisfied with (i)) to ensure a reasonably high order.

The result is a class of k -step formulas of order $k+2$, which are of the form

$$y_{n+1} = y_n + h \sum_{i=0}^k \beta_i f_{n+i-k+1} + h^2 \gamma_k g_{n+1}. \quad (3.7)$$

The first few of these methods are

$$\begin{aligned} k=1: \quad y_{n+1} &= y_n + h \left(\frac{2}{3} f_{n+1} + \frac{1}{3} f_n \right) - \frac{1}{6} h^2 g_{n+1} \\ k=2: \quad y_{n+1} &= y_n + h \left(\frac{29}{48} f_{n+1} + \frac{5}{12} f_n - \frac{1}{48} f_{n-1} \right) - \frac{1}{8} h^2 g_{n+1} \\ k=3: \quad y_{n+1} &= y_n + h \left(\frac{307}{540} f_{n+1} + \frac{19}{40} f_n - \frac{1}{20} f_{n-1} + \frac{7}{1080} f_{n-2} \right) \\ &\quad - \frac{19}{180} h^2 g_{n+1} \\ k=4: \quad y_{n+1} &= y_n + h \left(\frac{3133}{5760} f_{n+1} + \frac{47}{90} f_n - \frac{41}{480} f_{n-1} + \frac{1}{45} f_{n-2} \right. \\ &\quad \left. - \frac{17}{5760} f_{n-3} \right) - \frac{3}{32} h^2 g_{n+1} \end{aligned}$$

For a general expression, see Eq. (3.12) below and Exercise 1.

The stability analysis for second derivative methods is again done by linearizing and leads to

$$y' = \lambda y \quad \text{for which} \quad y'' = \lambda^2 y. \quad (3.8)$$

This, inserted into (3.3), gives as the characteristic equation

$$\sum_{i=0}^k (\alpha_i - \mu \beta_i - \mu^2 \gamma_i) \zeta^i = 0, \quad \mu = h\lambda \quad (3.9)$$

instead of (1.6). Equation (3.9) is, for $\zeta = e^{i\theta}$, a quadratic equation which gives rise to *two* root locus curves which, together, describe the stability domain. The Enright methods (3.7) turn out to be A -stable for $k = 1$ and 2 (hence for $p = 3$ and 4) and are stiffly stable for $k = 3, 4, 5, 6$ and 7. The corresponding values α (for $A(\alpha)$ -stability), D and the error constants C are given in Table 3.1. Pictures are shown in Fig. 3.1.

Table 3.1. Stability characteristics and error constants for Enright methods

k	1	2	3	4	5	6	7
p	3	4	5	6	7	8	9
α	90°	90°	87.88°	82.03°	73.10°	59.95°	37.61°
D	0.	0.	0.103	0.526	1.339	2.728	5.182
C	0.01389	0.00486	0.00236	0.00136	0.00086	0.00059	0.00042

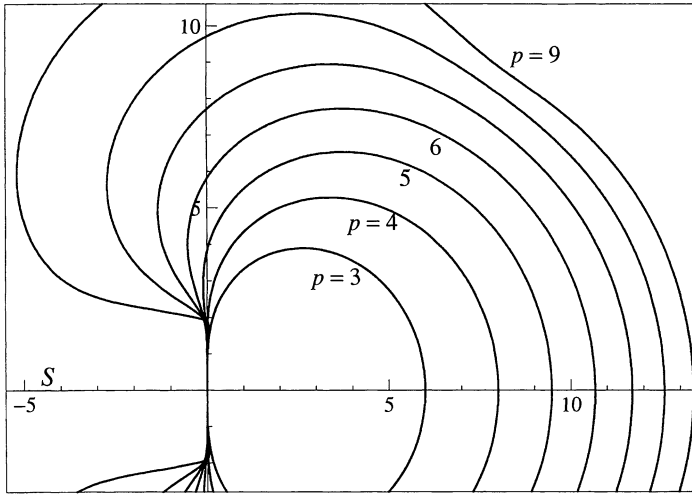


Fig. 3.1. Stability domains of Enright methods

Dense Output for Enright Methods. We have seen in Sect. III.1 that Newton's interpolation formula, based on the data $x_{n+1}, x_n, \dots, x_{n-k+1}$,

- when integrated from x_n to x_{n+1} , leads to the implicit Adams methods;
- when differentiated at x_{n+1} , leads to the BDF methods.

It is natural to apply the same idea to *Hermite* interpolation (Addison 1979): guided by much previous experience (see above) we choose the data points

$$x_{n+1} \text{ (double node), } x_n, x_{n-1}, \dots, x_{n-k+1} \text{ (simple nodes).} \quad (3.10)$$

This gives the following scheme of divided differences

$$\begin{array}{ccccccc}
 s = 1 & f_1 & & & & & \\
 & & hf'_1 & & & & \\
 s = 1 & f_1 & hf'_1 - \nabla f_1 & & & & \\
 & & \nabla f_1 & & \frac{hf'_1 - \nabla f_1 - \frac{1}{2}\nabla^2 f_1}{2!} & & \\
 s = 0 & f_0 & \frac{\nabla^2 f_1}{2!} & & & & \\
 & & \nabla f_0 & & & & \\
 s = -1 & f_{-1} & & & & &
 \end{array}$$

where $x = x_n + sh$. For these “confluent” data, Newton’s interpolation formula becomes

$$\begin{aligned}
 f(x_n + sh) = & f_1 + (s-1)hf'_1 + (s-1)^2(hf'_1 - \nabla f_1) \\
 & + (s-1)^2s \frac{hf'_1 - \nabla f_1 - \frac{1}{2}\nabla^2 f_1}{2!} \\
 & + (s-1)^2s(s+1) \frac{hf'_1 - \nabla f_1 - \frac{1}{2}\nabla^2 f_1 - \frac{1}{3}\nabla^3 f_1}{3!} + \dots
 \end{aligned} \quad (3.11)$$

We now interpret f as the derivative $f(x, y(x))$ of the solution, so that f' becomes the second derivative. Integrating Formula (3.11) from x_n to x_{n+1} we obtain

$$y_{n+1} = y_n + hf_{n+1} - h \sum_{j=1}^k \frac{\nabla^j f_{n+1}}{j} \left(\sum_{i=j}^k \nu_i \right) + h^2 g_{n+1} \cdot \left(\sum_{i=0}^k \nu_i \right) \quad (3.12)$$

where

$$\nu_i = \int_0^1 \frac{(s-1)^2 s(s+1) \dots (s+i-2)}{i!} ds = (-1)^i \int_0^1 (s-1) \binom{1-s}{i} ds. \quad (3.13)$$

Table 3.2. Coefficients for Enright methods

j	0	1	2	3	4	5	6	7
ν_j	$-\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{24}$	$\frac{7}{360}$	$\frac{17}{1440}$	$\frac{41}{5040}$	$\frac{731}{120960}$	$\frac{8563}{1814400}$

The first few values of ν_i are given in Table 3.2 and Eq. (3.12) is seen to be identical with (3.7). Dense output, of course, is obtained by integrating (3.11) from x_n to $x_n + \theta h$:

$$y(x_n + \theta h) \approx y_n + \theta h f_{n+1} - h \sum_{j=1}^k \frac{\nabla^j f_{n+1}}{j} \left(\sum_{i=j}^k \nu_i(\theta) \right) + h^2 g_{n+1} \cdot \left(\sum_{i=0}^k \nu_i(\theta) \right)$$

where

$$\nu_i(\theta) = (-1)^i \int_0^\theta (s-1) \binom{1-s}{i} ds.$$

Second Derivative BDF Methods

If we are interested in a “second derivative” analogue of the BDF methods, we replace all f ’s by y ’s in (3.11) and differentiate twice at x_{n+1} . This, on setting $y''(x_{n+1}) = g_{n+1}$, results in the methods

$$\frac{h^2}{2} g_{n+1} = \left(\sum_{i=1}^k \frac{1}{i} \right) h f_{n+1} - \sum_{j=1}^k \left(\sum_{i=j}^k \frac{1}{i} \right) \frac{\nabla^j y_{n+1}}{j} \quad (3.14)$$

which we call “*Second derivative BDF methods*” (SDBDF, the reader is cautioned against confusion: Cash (1981) uses this expression for the class of “Enright methods”). Analyzing the stability of these methods leads to the parameters of Table 3.3. The root locus curves are drawn in Fig. 3.2.

In complete analogy to the behaviour of implicit Adams compared to BDF methods, the second derivative BDF methods have larger error constants than the Enright methods, but allow stiffly stable methods of higher order.

Table 3.3. Stability characteristics and error constants for SDBDF methods

k	1	2	3	4	5	6	7	8	9	10
p	2	3	4	5	6	7	8	9	10	11
α	90°	90°	90°	89.36°	86.35°	80.82°	72.53°	60.71°	43.39°	12.34°
D	0.	0.	0.	0.015	0.128	0.401	0.886	1.646	2.770	4.373
C	.1667	.0556	.0273	.0160	.0104	.0073	.0054	.0041	.0032	.0026

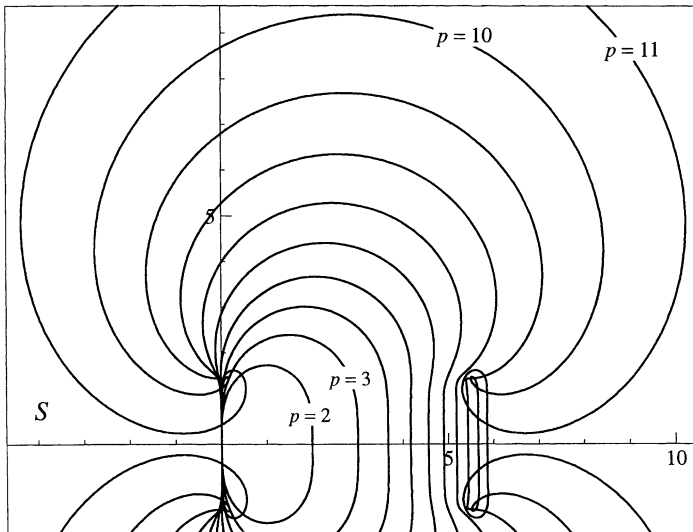


Fig. 3.2. Root locus curves of SDBDF methods

Blended Multistep Methods

The original motivation for *blended methods* goes as follows (Skeel & Kong 1977): We know that Adams methods

$$-y_{n+1} + y_n + h(\beta_k f_{n+1} + \beta_{k-1} f_n + \dots + \beta_0 f_{n-k+1}) = 0 \quad (AMF^{(k+1)})$$

are a very good choice for nonstiff problems, and that BDF methods

$$-(\alpha_k y_{n+1} + \alpha_{k-1} y_n + \dots + \alpha_0 y_{n-k+1}) + h f_{n+1} = 0 \quad (BDF^{(k)})$$

are a very good choice for stiff problems. Nonstiff problems are characterized by the fact that $-h \partial f / \partial y$ is *small*, while stiff problems are characterized by *large* $-h \partial f / \partial y$ (at first this makes sense only for scalar equations; but it works as well for systems of equations if we descend into the eigenspaces of the Jacobian matrix $\partial f / \partial y = J$). The idea is now to use a weighted mean (“blend”, a term suggested by C.W. Gear) of the two methods such as

$$\{AMF^{(k+1)}\} - \gamma^{(k)} h J \{BDF^{(k)}\} = 0 \quad (3.15)$$

where $\gamma^{(k)}$ is a free parameter. The factor $-hJ$, when small or large, just puts the weight at the right place, as required by the above motivation. Taylor expansion shows that Eq. (3.15) is for all $\gamma^{(k)}$ of order $p = k + 1$ (the factor “ h ” in the second term saves one order), even if J differs from $\partial f / \partial y$. This method is thus a multistep analogue to the W -methods discussed in Sect. IV.7.

Example. We put $k = 2$ in (3.15) and insert the values from Sect. III.1, Formulas (III.1.8”) and (III.1.22”):

$$y_{n+1} = y_n + h \left(\frac{5}{12} f_{n+1} + \frac{8}{12} f_n - \frac{1}{12} f_{n-1} \right) - \gamma^{(2)} h J \left(-\frac{3}{2} y_{n+1} + 2y_n - \frac{1}{2} y_{n-1} + h f_{n+1} \right). \quad (3.16)$$

If we now suppose that our differential equation is linear and autonomous $y' = Jy$, then $Jy_{n+i} = f_{n+i}$ and the equation simplifies. Two special choices for $\gamma^{(2)}$ are then interesting:

- $\gamma^{(2)} = 1/6$: In this case the f_{n-1} cancels with Jy_{n-1} and Eq. (3.16) becomes the $(k-1)$ -step Enright formula of order $k+1$;
- $\gamma^{(2)} = 1/8$: This is a “superconvergence point” for linear equations and we obtain the k -step Enright formula of order $k+2$.

Both properties generalize to arbitrary k ; in the first case we have to put $\gamma^{(k)} = -k\gamma_k^*$, where the γ_k^* are the values of Table III.1.2, and in the second case we use $\gamma^{(k)} = -\sum_{i=0}^k \nu_i$ as in (3.12). Blended methods therefore share the excellent stability properties of the Enright methods and seem, at the same time, easier to implement. A third possibility is to choose $\gamma^{(k)}$ in order to maximize the angle α

Table 3.4. Values for $\gamma^{(k)}$ and corresponding angles for blended methods

k	p	$-k\gamma_k^*$	α for $\gamma^{(k)} = -k\gamma_k^*$	$\gamma_{opt}^{(k)}$	α for $\gamma^{(k)} = \gamma_{opt}^{(k)}$
1	2	.5	90°	$[0, +\infty)$	90°
2	3	.1666667	90°	$[\cdot 125, +\infty)$	90°
3	4	.125	90°	$[\cdot 12189, \cdot 68379]$	90°
4	5	.1055556	87.88°	.1284997	89.42°
5	6	.09375	82.03°	.1087264	86.97°
6	7	.08561508	73.10°	.0962596	82.94°
7	8	.07957176	59.95°	.08754864	77.43°
8	9	.07485229	37.61°	.08105624	70.22°
9	10	.07103299	—	.07599875	60.68°
10	11	.06785850	—	.07192937	47.63°
11	12	.06516462	—	.06857226	28.68°

for $A(\alpha)$ -stability. The root-locus-curve equation for general $\gamma^{(k)}$ becomes

$$\mu^2 \cdot \gamma^{(k)} + \mu \left(- \sum_{j=0}^k \gamma_j^* (1 - e^{-i\theta})^j - \gamma^{(k)} \sum_{j=1}^k \frac{1}{j} (1 - e^{-i\theta})^j \right) + (1 - e^{-i\theta}) = 0.$$

Skeel & Kong (1977) have carefully computed the optimal $\gamma^{(k)}$ (see Table 3.4, the imprecise values for the “Enright column” have been corrected) and arrived thereby at stiffly stable methods up to order 12.

Extended Multistep Methods of Cash

The second possibility for circumventing Dahlquist’s barrier, instead of adding higher derivatives, is to add further stages, additional nodes, or off-step points. This leads into the huge desert (“A fable of K. Burrage”) of general linear methods which have been discussed in Sect. III.8. Pioneering results for stiff differential equations are the “composite multistep methods” of Sloate & Bickart (1973), Bickart & Rubin (1974), the “hybrid” methods of England (1982), and the “extended” BDF methods of Cash (1980). We shall present the basic ideas for the latter in some detail. In order to increase stability of the BDF methods, we extend them by adding a “super-future” point at x_{n+k+1}

$$\sum_{j=0}^k \alpha_j y_{n+j} = h \beta_k f_{n+k} + h \beta_{k+1} f_{n+k+1}, \quad (3.17)$$

where the coefficients are obtained by solving $\sum_j \alpha_j j^q = q \sum_j \beta_j j^{q-1}$ for $q = 0, 1, \dots, k+1$ with the normalization $\alpha_k = 1$. Formula (3.17) is then used as follows (see Fig. 3.3):

- (i) Suppose that the solution values $y_n, y_{n+1}, \dots, y_{n+k-1}$ are available. Compute \bar{y}_{n+k} as the solution of the conventional BDF formula

$$\sum_{j=0}^k \hat{\alpha}_j y_{n+j} = h \hat{\beta}_k f_{n+k}, \quad \hat{\alpha}_k = 1; \quad (3.17i)$$

- (ii) Compute \bar{y}_{n+k+1} as the solution of the same BDF formula advanced by one step (using \bar{y}_{n+k} for y_{n+k})

$$\sum_{j=0}^k \hat{\alpha}_j y_{n+j+1} = h \hat{\beta}_k f_{n+k+1} \quad (y_{n+k} := \bar{y}_{n+k}) \quad (3.17ii)$$

and set $\bar{f}_{n+k+1} = f(x_{n+k+1}, \bar{y}_{n+k+1})$;

- (iii) Discard \bar{y}_{n+k} , insert \bar{f}_{n+k+1} into (3.17) and solve for a new y_{n+k} which serves as the final numerical solution of the method.

The advance of the numerical solution by *one step* thus requires the solution of *three* nonlinear systems of dimension n . In stage (i) and stage (iii) we have excellent initial approximations: the super future point of the previous step and the value \bar{y}_{n+k} , respectively.

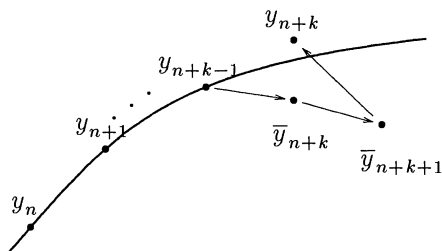


Fig. 3.3. Errors of Cash's algorithm

Lemma 3.1 (Cash 1980). *If Formula (3.17) is of order $k+1$ and the BDF methods used in (3.17i) and (3.17ii) are of order k , then the whole predictor-corrector algorithm (i)–(iii) is of order $k+1$.*

Proof. Suppose that y_n, \dots, y_{n+k-1} are on the exact solution (Fig. 3.3). Then a simple calculation (as in the proof of Lemma III.2.2, see also Eq. (III.2.7)) shows that

$$y(x_{n+k}) - \bar{y}_{n+k} = C_1 h^{k+1} y^{(k+1)}(x_{n+k}) + \mathcal{O}(h^{k+2}) \quad (3.18)$$

$$y(x_{n+k+1}) - \bar{y}_{n+k+1} = C_1 \left(1 - \frac{\hat{\alpha}_{k-1}}{\hat{\alpha}_k}\right) h^{k+1} y^{(k+1)}(x_{n+k}) + \mathcal{O}(h^{k+2}) \quad (3.19)$$

where C_1 depends on the BDF method used. If now $C_2 h^{k+2} y^{(k+2)}(\xi)$ is the defect of Eq. (3.17) (for the exact solution), replacing $h f(x_{n+k+1}, y(x_{n+k+1}))$ by

$hf(x_{n+k+1}, \bar{y}_{n+k+1})$ adds the expression obtained in (3.19) to this error and we obtain

$$y(x_{n+k}) - y_{n+k} = h^{k+2} \left(C_2 y^{(k+2)} + \beta_{k+1} C_1 \left(1 - \frac{\hat{\alpha}_{k-1}}{\alpha_k} \right) \frac{\partial f}{\partial y} \cdot y^{(k+1)} \right) (x_{n+k}) + \mathcal{O}(h^{k+3}). \quad (3.20)$$

The method is thus of order $k+1$. Like Runge-Kutta methods, but unlike linear multistep methods, the principal error term is composed of several “elementary differentials”. \square

Modified EBDP Methods. A disadvantage of the above algorithm is that stages (i) and (ii) represent nonlinear systems with the same Jacobian $I - h\hat{\beta}_k J$, but stage (iii) has a different Jacobian $I - h\beta_k J$. This requires an extra LU-decomposition. The idea is to modify Eq. (3.17) for stage (iii) as follows (Cash 1983):

$$\sum_{j=0}^k \alpha_j y_{n+j} = h\hat{\beta}_k f_{n+k} + h(\beta_k - \hat{\beta}_k) \bar{f}_{n+k} + h\beta_{k+1} \bar{f}_{n+k+1}. \quad (3.17.\text{mod})$$

This just adds an extra h^{k+2} -term to the above proof and does not alter the order of the method. It allows the same Jacobian to be used in the Newton iteration for all three stages, and, possibly, to preserve it over several steps as well.

Stability Analysis. We insert $hf_j = \mu y_j$ in (3.17.mod), (3.17i) and (3.17ii), set $y_n = 1$, $y_{n+1} = \zeta$, \dots , $y_{n+k-1} = \zeta^{k-1}$ and compute, following the algorithm (i), (ii), (iii), the solution $y_{n+k} =: \zeta^k$. This gives the characteristic equation

$$A\mu^3 + B\mu^2 + C\mu + D = 0 \quad (3.21)$$

where

$$\begin{aligned} A &= \hat{\beta}_k^3 \zeta^k \\ B &= -2\hat{\beta}_k^2 \zeta^k + \hat{\beta}_k(\beta_k - \hat{\beta}_k)R + \hat{\beta}_k \beta_{k+1} S - \hat{\beta}_k^2 T \\ C &= \hat{\beta}_k \zeta^k + (\hat{\alpha}_{k-1} \beta_{k+1} - \beta_k + \hat{\beta}_k)R - \beta_{k+1} S + 2\hat{\beta}_k T \\ D &= -T \\ R &= \sum_{j=0}^{k-1} \hat{\alpha}_j \zeta^j, \quad S = \sum_{j=0}^{k-2} \hat{\alpha}_j \zeta^{j+1}, \quad T = \sum_{j=0}^k \alpha_j \zeta^j. \end{aligned} \quad (3.22)$$

Inserting $\zeta = e^{i\theta}$, Equation (3.21) gives us three roots $\mu_i(\theta)$ $i = 1, 2, 3$, which describe the stability domain. These, computed by Cardano’s formula, are displayed in Fig. 3.4. The corresponding stability characteristics are given in Table 3.5. The methods are A -stable for $p \leq 4$ and are stiffly stable for orders up to 9.

Table 3.5. Stability measures for Cash’s modified EBDf methods

k	1	2	3	4	5	6	7	8
p	2	3	4	5	6	7	8	9
α	90°	90°	90°	88.36°	83.07°	74.48°	61.98°	42.87°
D	0.	0.	0.	0.040	0.246	0.684	1.402	2.432

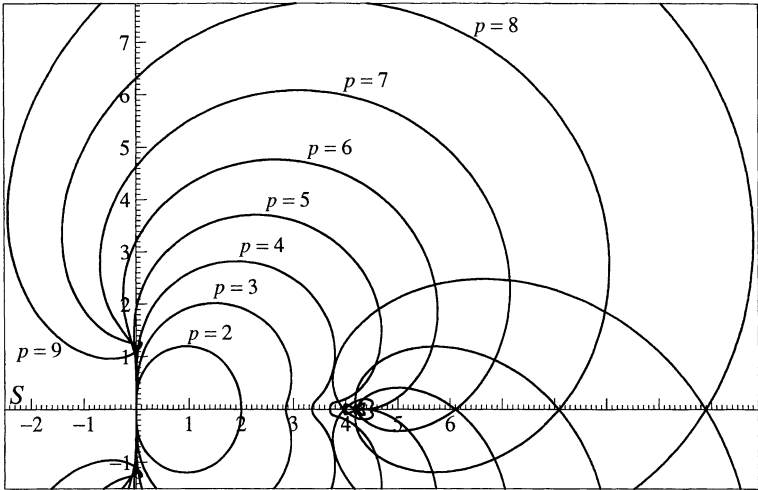


Fig. 3.4. Stability domains for Cash’s MEBDF methods

Multistep Collocation Methods

... a theorem of great antiquity ... the simple theorem of polynomial interpolation upon which much practical numerical analysis rests ...
(P.J. Davis, Interp. and Approx., Chapter II, 1963)

There are essentially two possibilities to extend the idea of collocation, which is so successful in the Runge-Kutta case (see Sect. II.7, Formulas (II.7.16)), into the multistep scene:

- a) In a Nordsieck type manner with given $y_n, hy'_n, h^2y''_n/2, \dots$ compute $y_{n+1}, hy'_{n+1}, h^2y''_{n+1}/2, \dots$. The result is a spline function which approximates the solution globally. Butcher’s generalized singly-implicit methods (Butcher 1981) are of this type. Extensive studies of these methods are due to Mülthei (1982).
- b) In a multistep manner with given $y_n, y_{n-1}, \dots, y_{n-k+1}$ compute y_{n+1} , then discard, as usual, the last point y_{n-k+1} and continue. This possibility was first proposed and analysed by Guillou & Soulé (1969). It is also the subject of a paper by Lie & Nørsett (1989) and will retain our attention here in more detail. In evident generalization of Definition II.7.6, the method is defined as follows:

Definition 3.2. Let s real numbers c_1, \dots, c_s (typically between 0 and 1) be given and k solution values $y_n, y_{n-1}, \dots, y_{n-k+1}$. Then define the corresponding *collocation polynomial* $u(x)$ of degree $s + k - 1$ by (see Fig. 3.5)

$$u(x_j) = y_j \quad j = n - k + 1, \dots, n \quad (3.23a)$$

$$u'(x_n + c_i h) = f(x_n + c_i h, u(x_n + c_i h)) \quad i = 1, \dots, s. \quad (3.23b)$$

The numerical solution is then

$$y_{n+1} := u(x_{n+1}). \quad (3.23c)$$

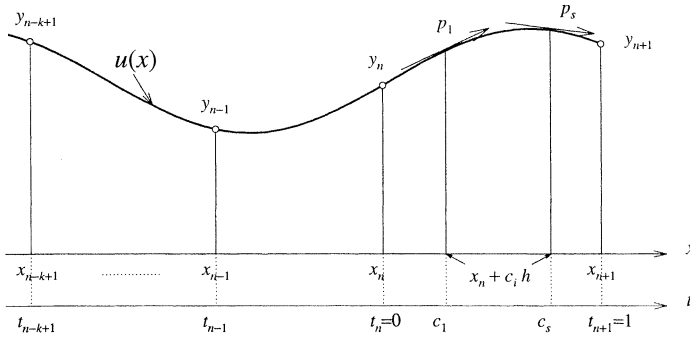


Fig. 3.5. The collocation polynomial

If we suppose the derivatives $u'(x_n + c_i h)$ are known, Eqs. (3.23a) and (3.23b) constitute a Hermite interpolation problem with incomplete data: the function values at $x_n + c_i h$ are missing. We therefore have no nice formulas and reduce the problem to a linear algebraic equation. We introduce the dimensionless coordinate $t = (x - x_n)/h$, $x = x_n + th$, nodes $t_1 = -k + 1, \dots, t_{k-1} = -1$, $t_k = 0$ and define polynomials $\varphi_i(t)$ ($i = 1, \dots, k$) of degree $s + k - 1$ by

$$\begin{aligned} \varphi_i(t_j) &= \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \quad j = 1, \dots, k \\ \varphi'_i(c_j) &= 0 \quad j = 1, \dots, s \end{aligned} \quad (3.24)$$

and polynomials $\psi_i(t)$ ($i = 1, \dots, s$) by

$$\begin{aligned} \psi_i(t_j) &= 0 \quad j = 1, \dots, k \\ \psi'_i(c_j) &= \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad j = 1, \dots, s. \end{aligned} \quad (3.25)$$

This makes these polynomials a (generalized) Lagrange basis and the polynomial $u(x)$ is readily written as

$$u(x_n + th) = \sum_{j=1}^k \varphi_j(t) y_{n-k+j} + h \sum_{j=1}^s \psi_j(t) u'(x_n + c_j h). \quad (3.26)$$

Formulas (3.24) and (3.25) do not always have a solution (Exercise 4 below). A convenient way of computing them is indicated in Exercise 5. Putting $t = c_i$ in (3.26), writing $u(x_n + c_i h) = v_i$ and inserting the collocation condition (3.23b) we obtain

$$v_i = \sum_{j=1}^k \varphi_j(c_i) y_{n-k+j} + h \sum_{j=1}^s \psi_j(c_i) f(x_n + c_j h, v_j) \quad (3.27a)$$

$$i = 1, \dots, s$$

$$y_{n+1} = \sum_{j=1}^k \varphi_j(1) y_{n-k+j} + h \sum_{j=1}^s \psi_j(1) f(x_n + c_j h, v_j), \quad (3.27b)$$

a general linear method as defined in (III.8.7).

Theorem 3.3. *The collocation method (3.23) is equivalent to the general linear method*

$$v_i = \sum_{j=1}^k a_{ij} y_{n-k+j} + h \sum_{j=1}^s b_{ij} f(x_n + c_j h, v_j) \quad i = 1, \dots, s \quad (3.28)$$

$$y_{n+1} = \sum_{j=1}^k a_{k+1,j} y_{n-k+j} + h \sum_{j=1}^s b_{k+1,j} f(x_n + c_j h, v_j)$$

where

$$a_{ij} = \varphi_j(c_i), \quad b_{ij} = \psi_j(c_i), \quad a_{k+1,j} = \varphi_j(1), \quad b_{k+1,j} = \psi_j(1) \quad (3.29)$$

and $\varphi_j(t)$, $\psi_j(t)$ are polynomials defined by (3.24) and (3.25). Formula (3.26) provides a continuous output. \square

A straightforward extension of the proof of Theorem II.7.9, again using the Gröbner & Alekseev formula (I.14.18), yields

Theorem 3.4 (Guillou & Soulé 1969). *If the quadrature formula (3.27b) is exact for polynomials $g(t)$ of degree $\leq s + k + r$, i.e., $\sum_{j=1}^k \varphi_j(1) = 1$ and*

$$\sum_{j=1}^k \varphi_j(1) \int_{j-k}^1 g(t) dt = \sum_{i=1}^s \psi_i(1) g(c_i),$$

then the multistep collocation method (3.28) also has order $s + k + r$. \square

Methods of “Radau” Type

Nous allons maintenant étudier une classe de formules qui généralise les formules ordinaires de Gauss, Radau et Lobatto.
(Guillou & Soulé 1969)

An interesting question is now how to choose the nodes c_i in order to obtain the highest possible order. Using an elegant idea of Krylov (1959) (see the last chapter of his book on integration), Guillou & Soulé (1969) and Lie & Nørsett (1989) constructed such methods of maximal order $p = 2s + k - 1$. Unfortunately, these methods are not stiffly stable and therefore of no use for stiff problems. Consequently, we fix $c_s = 1$ to achieve stability at infinity and try to determine c_1, \dots, c_{s-1} so that the order becomes $p = 2s + k - 2$. Because of Theorem 3.4, it is sufficient to consider quadrature problems.

And now to Krylov's idea for integrals, adapted to our situation. We fill in the gaps in the data for Hermite interpolation, i.e., we suppose that the *function values* $v_i = u(x_n + c_i h)$ ($i = 1, \dots, s-1$) are known and we extend our Lagrange basis accordingly: firstly, we add polynomials $\chi_1(t), \dots, \chi_{s-1}(t)$ of degree $2s + k - 2$ which must satisfy

$$\chi_i(t_j) = 0 \quad j = 1, \dots, k \quad (3.30a)$$

$$\chi'_i(c_j) = 0 \quad j = 1, \dots, s \quad (3.30b)$$

$$\chi_i(c_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad j = 1, \dots, s-1 \quad (3.30c)$$

(Caution: the last condition is *not* for $j = s$, because c_s is not a free node). Secondly, the polynomials $\varphi_i(t)$ and $\psi_i(t)$ are replaced by $\tilde{\varphi}_i(t)$, $\tilde{\psi}_i(t)$ of degree $2s + k - 2$ which, in addition to (3.24) and (3.25), must satisfy

$$\tilde{\varphi}_i(c_j) = 0 \quad \text{and} \quad \tilde{\psi}_i(c_j) = 0 \quad j = 1, \dots, s-1. \quad (3.31)$$

Then Eq. (3.26) is replaced by

$$\tilde{u}(x_n + th) = \sum_{j=1}^k \tilde{\varphi}_j(t) y_{n-k+j} + \sum_{j=1}^{s-1} \chi_j(t) v_j + h \sum_{j=1}^s \tilde{\psi}_j(t) u'(x_n + c_j h), \quad (3.32)$$

and (3.27b) becomes the integration formula

$$y_{n+1} = \sum_{j=1}^k \tilde{\varphi}_j(1) y_{n-k+j} + \sum_{j=1}^{s-1} \chi_j(1) v_j + h \sum_{j=1}^s \tilde{\psi}_j(1) u'(x_n + c_j h) \quad (3.33)$$

which is of order $2s + k - 2$. If now, by a miracle, all coefficients

$$\chi_j(1) = 0 \quad (j = 1, \dots, s-1) \quad (3.34)$$

were zero, then the quadrature Formula (3.27b) would become equal to (3.33), since by uniqueness the remaining coefficients $\tilde{\varphi}_j(1)$ and $\tilde{\psi}_j(1)$ must also be equal to $\varphi_j(1)$ and $\psi_j(1)$.

Theorem 3.5. *If the collocation points c_1, \dots, c_{s-1} (with $c_s = 1$) are chosen such that the polynomials $\varphi_i(t), \psi_i(t)$ of (3.24), (3.25) exist uniquely and that (3.34) is true, then the collocation method (3.28) is of highest possible order $2s + k - 2$. \square*

Computation of the Nodes. Equation (3.34) together with the conditions (3.30) allow us to write the polynomials $\chi_i(t)$ in the simple form

$$\chi_i(t) = C \prod_{j=1}^k (t - t_j) \prod_{\substack{j=1 \\ j \neq i}}^s (t - c_j)^2. \quad (3.35)$$

where C is determined by $\chi_i(c_i) = 1$. This then satisfies all derivative requirements (3.30b), except at c_i . $\chi'_i(c_i)$ is readily computed from (3.35) by taking logarithms and the conditions $\chi'_i(c_i) = 0$ give

$$\sum_{j=1}^k \frac{1}{c_i - t_j} + \sum_{\substack{j=1 \\ j \neq i}}^s \frac{2}{c_i - c_j} = 0, \quad i = 1, \dots, s-1. \quad (3.36)$$

Example. For the case $s = 3$, Eqs. (3.36) become ($c_3 = 1$)

$$\begin{aligned} \frac{2}{c_2 - c_1} &= \frac{2}{c_1 - 1} + \sum_{j=1}^k \frac{1}{c_1 - t_j}, \\ \frac{2}{c_1 - c_2} &= \frac{2}{c_2 - 1} + \sum_{j=1}^k \frac{1}{c_2 - t_j}. \end{aligned} \quad (3.37)$$

These two equations can easily be solved for c_2 and c_1 respectively, and lead to the curves displayed for $k = 3$ and $k = 4$ in Fig. 3.6. We see that a huge number of solutions is possible (precisely $\binom{s+k-1}{k-1}$, Krylov imagined charged electrical particles in equilibrium to prove their existence), but most of these lead to totally unstable and therefore useless methods (in the sense of Sect. III.3). Thus the only choice which we retain are the rightmost solutions c_i with $0 < c_1, c_2 < 1$, shown in Table 3.6 below. In addition, as Krylov has shown (see Krylov (1959), English translation 1962, p. 329) this choice leads to the smallest error constant (for once, stability and small error are *not* in conflict!)

Stability of the Radau-Type Methods. The stability analysis of the Radau methods is done by inserting $y' = \lambda y$ into (3.28). Since $c_s = 1$ we have $y_{n+1} = v_s$ and thus obtain (for $s = 3$) the characteristic equation

$$\begin{pmatrix} 1 - \mu b_{11} & -\mu b_{12} & -\mu b_{13} \\ -\mu b_{21} & 1 - \mu b_{22} & -\mu b_{23} \\ -\mu b_{31} & -\mu b_{32} & 1 - \mu b_{33} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \zeta^3 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} 1 \\ \zeta \\ \zeta^2 \end{pmatrix},$$

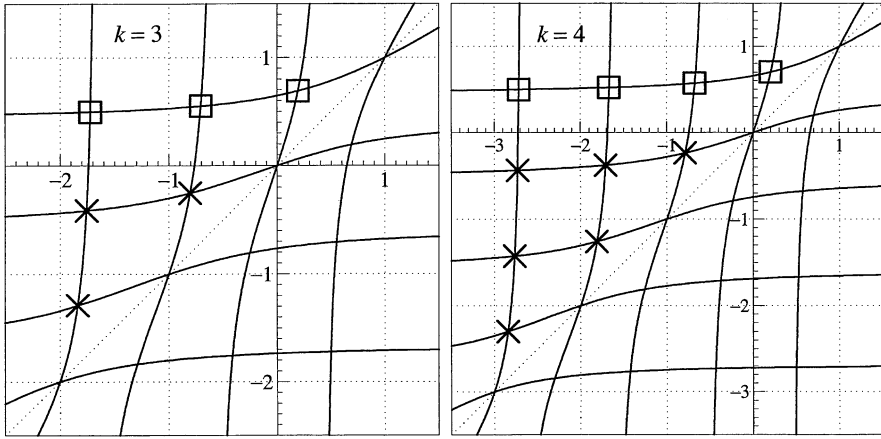


Fig. 3.6. Solutions of (3.37). \times unstable, \square stable

or

$$\zeta^3 = (0, 0, 1) \begin{pmatrix} 1 - \mu b_{11} & -\mu b_{12} & -\mu b_{13} \\ -\mu b_{21} & 1 - \mu b_{22} & -\mu b_{23} \\ -\mu b_{31} & -\mu b_{32} & 1 - \mu b_{33} \end{pmatrix}^{-1} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} 1 \\ \zeta \\ \zeta^2 \end{pmatrix} \quad (3.38)$$

which, when multiplied by $\det(I - \mu B)$, becomes a polynomial of degree 3 in μ . For a general multistep collocation method (3.28) we obtain in this way

$$q_k(\mu)\zeta^k + q_{k-1}(\mu)\zeta^{k-1} + \dots + q_0(\mu) = 0$$

where $q_k(\mu) = \det(I - \mu B)$ and all $q_i(\mu)$ are polynomials of degree at most s .

The root locus curves of Fig. 3.7 were again obtained by Cardano's formula. Coefficients and stability measures are given in Table 3.6. The methods for $k = 1, 2$ (orders $p = 5$ and 6) are A -stable. The subsequent methods have surprisingly large α -values for very high orders (up to $p \approx 20$), which makes this class very promising.

Exercises

1. Show that the coefficients ν_j in (3.13) for the Enright methods can be computed recursively by

$$\nu_j = -\frac{1}{(j+1)(j+2)} - \sum_{k=0}^{j-1} \nu_k S_{j+1-k} \quad \text{where} \quad S_l = \sum_{k=1}^l \frac{1}{k(l+1-k)}.$$

Hint. See the proof of Eq. (III.1.7). The generating function $G(t) = \sum_{j=0}^{\infty} \nu_j t^j$ becomes here $\int_0^1 (s-1)(1-t)^{1-s} ds$.

Table 3.6. Coefficients and stability measures
for multistep Radau methods ($s = 3$)

k	p	c_1	c_2	c_3	α	D
1	5	0.155051025721682	0.644948974278318	1.	90°	0.000
2	6	0.177891722985607	0.673235257220651	1.	90°	0.000
3	7	0.192169638937766	0.689317969824851	1.	89.73°	0.016
4	8	0.202814874040288	0.700407719104611	1.	89.13°	0.084
5	9	0.211395456069620	0.708798418188500	1.	88.61°	0.178
6	10	0.218626151232186	0.715507419158199	1.	88.14°	0.278
7	11	0.224897548200883	0.721072684914921	1.	87.70°	0.376
8	12	0.230448266933707	0.725812172023161	1.	87.28°	0.467
9	13	0.235435607740434	0.729928926504599	1.	86.89°	0.555
10	14	0.239969169367303	0.733560240031675	1.	86.51°	0.649
11	15	0.244128606044551	0.736803122952198	1.	86.14°	0.763
12	16	0.247973766491964	0.739728565298052	1.	85.79°	0.917
13	17	0.251550844436705	0.742390019356757	1.	85.44°	1.135
14	18	0.254896295040291	0.744828697795402	1.	85.07°	1.462
15	19	0.258039429919700	0.747077018862741	1.	84.68°	1.995
16	20	0.261004194709515	0.749160923778290	1.	84.23°	3.037

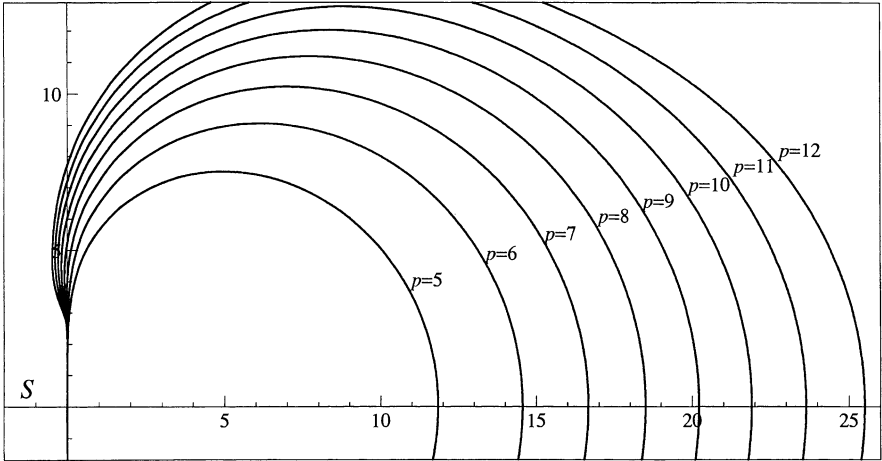


Fig. 3.7. Root locus curves for multistep Radau methods ($s = 3$)

2. The Enright Formulas are stiffly stable for $k \leq 7$ and are *not* stiffly stable, as one can easily inspect, e.g. by a computer plot, for $k = 8$, $k = 9, \dots$ and so on. Hence, everybody agrees that they are not stiffly stable for *any* $k > 7$. However, no rigorous proof has been found for this, as for instance the proof of Theorem III.3.4. Why don't you try to find one?

3. Prove that the second derivative BDF methods (3.14) are unstable (in the sense of Sect. III.3) for $k > 11$.
4. a) Show that for $k = 2$, $t_1 = -1$, $t_2 = 0$, $s = 1$, $c_1 = -1/2$ neither equations (3.24) nor equations (3.25) possess a solution.
- b) Show that (3.24) and (3.25) always admit unique solutions if all c_i are distinct and satisfy $c_i \geq 0$.
- Hint for (b).* If φ_i (or ψ_i) are written as $\sum_{l=1}^{s+k} a_l t^{l-1}$, then (3.24) and (3.25) become linear systems with the same matrix and different right-hand sides. The corresponding *homogeneous* system then possesses a non-zero solution iff the interpolation problem

$$\begin{aligned} p(t_j) &= 0 & j &= 1, \dots, k \\ p'(c_j) &= 0 & j &= 1, \dots, s \end{aligned} \quad (3.39)$$

has a non-zero solution. Since $p'(t)$ has at most $k + s - 2$ real zeros and since (Rolle's theorem) each interval (t_l, t_{l+1}) must contain at least one of these, there can be at most $s - 1$ zeros beyond $t_k = 0$.

5. A convenient way of computing the polynomials (3.24), (3.25) (written here for the case $s = 3$) is to put

$$\varphi_i(t) = (a_1 + a_2 t + a_3 t^2 + a_4 t^3) \prod_{l=1, l \neq i}^k (t - t_l). \quad (3.40)$$

Show that Eqs. (3.24) (for $i = j$) and (3.25) then become the following linear system

$$\begin{aligned} a_1 + t_i a_2 + t_i^2 a_3 + t_i^3 a_4 &= 1/r_i, \\ s_j a_1 + (s_j c_j + 1) a_2 + (s_j c_j^2 + 2c_j) a_3 + (s_j c_j^3 + 3c_j^2) a_4 &= 0, \quad j = 1, 2, 3 \end{aligned} \quad (3.41)$$

where $r_i = \prod_{l=1, l \neq i}^k (t_i - t_l)$, $s_j = \sum_{l=1, l \neq i}^k \frac{1}{c_j - t_l}$. Secondly, for

$$\psi_i(t) = (a_1 + a_2 t + a_3 t^2) \prod_{l=1}^k (t - t_l) \quad (3.42)$$

Eq. (3.25) becomes

$$s_j a_1 + (s_j c_j + 1) a_2 + (s_j c_j^2 + 2c_j) a_3 = \begin{cases} 0 & \text{if } j \neq i \\ 1/r_i & \text{if } j = i \end{cases} \quad j = 1, 2, 3$$

where $r_i = \prod_{l=1}^k (c_i - t_l)$, $s_j = \sum_{l=1}^k \frac{1}{c_j - t_l}$.

6. Generalize the proof and the result of Theorem IV.3.10 to multistep collocation methods.

Hint. Instead of $KM(x)$ in (IV.3.26) we have to insert a linear combination $\sum_{\ell=1}^k \alpha_\ell M_\ell(x)$ where $M_\ell(x) = M(x) \cdot x^{\ell-1}$, $M(x) = \frac{1}{s!} \prod_{i=1}^s (x - c_i)$ and $\alpha_1, \dots, \alpha_k$ are arbitrary. Instead of (IV.3.27) we then obtain

$$u(x) = \sum_{\ell=1}^k \alpha_\ell \sum_{j=0}^s \frac{M_\ell^{(j)}(x)}{\mu^j}. \quad (3.43)$$

Putting $x = t_1, t_2, \dots, t_k, t_{k+1}$ and $u(t_i) = y_i$ gives an overdetermined system for $\alpha_1, \dots, \alpha_k$ which has a solution only if a certain determinant is zero. Setting $y_1 = 1, y_2 = \zeta, y_3 = \zeta^2, \dots$ there leads to the characteristic equation

$$\det \begin{pmatrix} \sum_{j=0}^s M_1^{(j)}(t_1) \mu^{s-j} & \dots & \sum_{j=0}^s M_k^{(j)}(t_1) \mu^{s-j} & 1 \\ \sum_{j=0}^s M_1^{(j)}(t_2) \mu^{s-j} & \dots & \sum_{j=0}^s M_k^{(j)}(t_2) \mu^{s-j} & \zeta \\ \vdots & & \vdots & \vdots \\ \sum_{j=0}^s M_1^{(j)}(t_{k+1}) \mu^{s-j} & \dots & \sum_{j=0}^s M_k^{(j)}(t_{k+1}) \mu^{s-j} & \zeta^k \end{pmatrix}$$

as a generalization of (IV.3.22,23). Tedious expansions of this determinant into powers of ζ and μ (with many coefficients equal to zero) then leads to an explicit expression (see Theorem 7 of Lie 1990).

7. Prove that the 2-step 2-stage collocation method with $c_2 = 1$ is A -stable iff

$$c_1 \geq (\sqrt{17} - 1)/8.$$

Hint. a) Show that the characteristic equation is $q_2(\mu)\zeta^2 + q_1(\mu)\zeta + q_0(\mu) = 0$, where

$$\begin{aligned} q_2(\mu) &= -(9c_1 + 5) + \mu(3c_1^2 + 7c_1 + 2) - \mu^2 2c_1(c_1 + 1) \\ q_1(\mu) &= 12c_1 + 4 - \mu 4(c_1^2 - 1) \\ q_0(\mu) &= -3c_1 + 1 + \mu c_1(c_1 - 1). \end{aligned} \quad (3.44)$$

b) Apply Schur's criterion (1918) to the polynomial (3.44) with $\mu = it, t \in \mathbb{R}$.

Schur's criterion. Let $a(\zeta) = a_k \zeta^k + a_{k-1} \zeta^{k-1} + \dots + a_0$ ($a_k \neq 0$) be a polynomial with complex coefficients and set

$$a^*(\zeta) = \bar{a}_0 \zeta^k + \bar{a}_1 \zeta^{k-1} + \dots + \bar{a}_k.$$

Then, all zeros of $a(\zeta)$ lie inside the unit circle, iff

- i) $|a_0| < |a_k|$
- ii) the zeros of $\zeta^{-1}(a^*(0)a(\zeta) - a(0)a^*(\zeta))$, a polynomial of degree $k-1$, are all inside the unit circle.

8. Prove that $c_1 = (\sqrt{17} - 1)/8$ is a super-convergence point for the 2-step 2-stage collocation methods with $c_2 = 1$.

V.4 Order Stars on Riemann Surfaces

Riemann ist der Mann der glänzenden Intuition. Durch seine umfassende Genialität überragt er alle seine Zeitgenossen . . . Im Auftreten schüchtern, ja ungeschickt, musste sich der junge Dozent, zu dem wir Nachgeborenen wie zu einem Heiligen aufblicken, mancherlei Neckereien von seinen Kollegen gefallen lassen.

(F. Klein, Entwicklung der Mathematik im 19. Jhd., p. 246, 247)

We have seen in the foregoing sections that the highest possible order of A -stable linear multistep methods is two; furthermore, the second derivative Enright methods as well as the SDBDF methods were seen to be A -stable for $p \leq 4$; the three-stage Radau multistep methods were A -stable for $p \leq 6$. In this section we shall see that these observations are special cases of a general principle, the so-called “Daniel-Moore conjecture” which says that the order of an A -stable multistep method involving either s derivatives or s implicit stages satisfies $p \leq 2s$. Before proceeding to its proof, we should become familiar with Riemann surfaces.

Riemann Surfaces

Für manche Untersuchungen, namentlich für die Untersuchung algebraischer und Abel’scher Functionen ist es vortheilhaft, die Verzweigungsart einer mehrwerthigen Function in folgender Weise geometrisch darzustellen. Man denke sich in der (x, y) -Ebene eine andere mit ihr zusammenfallende Fläche (oder auf der Ebene einen unendlich dünnen Körper) ausgebreitet, welche sich so weit und nur so weit erstreckt, als die Function gegeben ist. Bei Fortsetzung dieser Function wird also diese Fläche ebenfalls weiter ausgedehnt werden. In einem Theile der Ebene, für welchen zwei oder mehrere Fortsetzungen der Function vorhanden sind, wird die Fläche doppelt oder mehrfach sein; sie wird dort aus zwei oder mehreren Blättern bestehen, deren jedes einen Zweig der Function vertritt. Um einen Verzweigungspunkt der Function herum wird sich ein Blatt der Fläche in ein anderes fortsetzen, so dass in der Umgebung eines solchen Punktes die Fläche als eine Schraubenfläche mit einer in diesem Punkte auf der (x, y) -Ebene senkrechten Axe und unendlich kleiner Höhe des Schraubenganges betrachtet werden kann. Wenn die Function nach mehreren Umläufen des z um den Verzweigungswerth ihren vorigen Werth wieder erhält (wie z.B. $(z - a)^{m/n}$, wenn m, n relative Primzahlen sind, nach n Umläufen von z um a), muss man dann freilich annehmen, dass sich das oberste Blatt der Fläche durch die übrigen hindurch in das unterste fortsetzt.

Die mehrwerthige Function hat für jeden Punkt einer solchen ihre Verzweigungsart darstellenden Fläche nur *einen* bestimmten Werth und kann daher als eine völlig bestimmte Function des Orts in dieser Fläche angesehen werden.

(B. Riemann 1857)

We take as example the BDF method (III.1.22'') for $k = 2$ which has the characteristic equation

$$\left(\frac{3}{2} - \mu\right)\zeta^2 - 2\zeta + \frac{1}{2}\mu = 0. \quad (4.1)$$

This quadratic equation expresses ζ as a function of μ , both are complex variables. It is immediately solved to yield

$$\zeta_{1,2} = \frac{2 \pm \sqrt{1+2\mu}}{3-2\mu} \quad (4.2)$$

which defines a *two-valued function*, i.e., to each $\mu \in \mathbb{C}$ we have two solutions ζ . These two solutions are displayed in Fig. 4.1 (we have plotted the level curves of $|\zeta_{1,2}(\mu)|$; the region with $|\zeta_1(\mu)| > 1$ is in white).

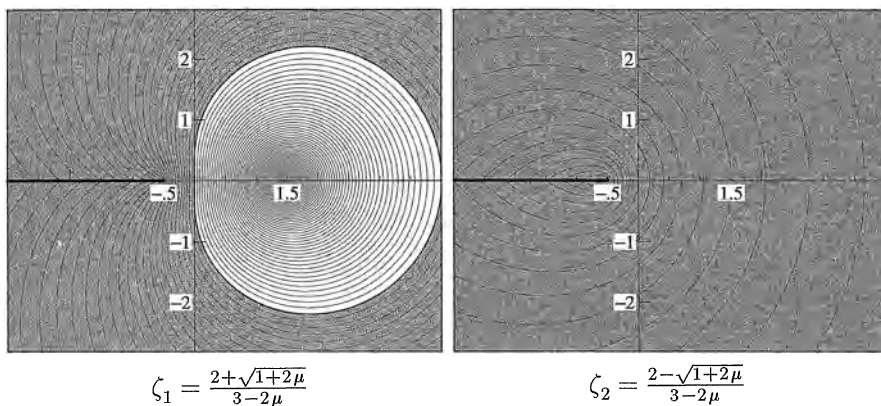


Fig. 4.1. The two solutions of the BDF2 characteristic equation

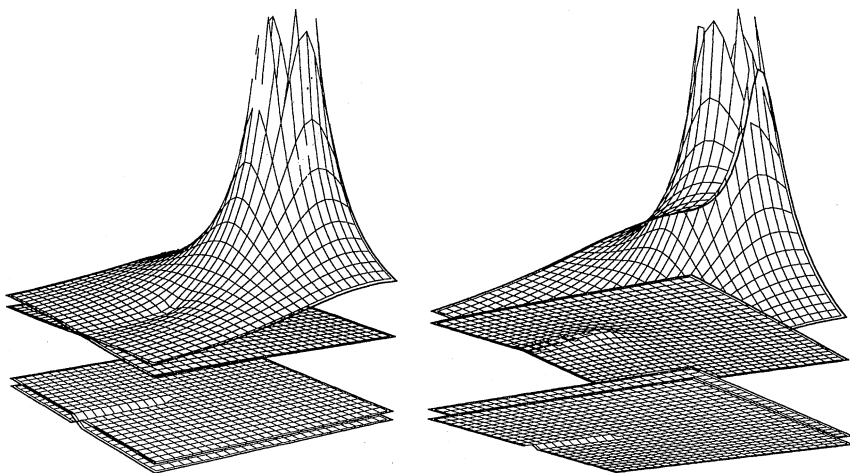


Fig. 4.2. Three dimensional view of the map (4.4)

We observe two essential facts. First, there is a pole of ζ_1 , but not of ζ_2 , at the point $\mu = 3/2$. This is due to the factor $(3/2 - \mu)$ in (4.1) which represents the implicit stage of the method. Second, we observe a curious discontinuity on the negative real axis left of the point $-1/2$, a phenomenon first observed in a famous paper of Puiseux (1850) (“... a encore cet inconvénient, que u devient alors une fonction discontinue ...”). It has its reason in the complex square root $\sqrt{1+2\mu}$ which, while $1+2\mu$ performs a revolution around the origin, only does *half* a revolution and exchanges the two roots. We cannot therefore speak in a natural way of the *two* complex functions $\zeta_1(\mu)$ and $\zeta_2(\mu)$. And here comes the great idea of Riemann (1857): Instead of varying μ in the complex plane \mathbb{C} , we imagine it varying in a *double sheet* of (in Riemann’s words: infinitely close) complex planes $\mathbb{C} \cup \mathbb{C}$. The μ ’s in the upper sheet are mapped to ζ_1 , the μ ’s in the lower sheet are mapped to ζ_2 . The double-valued function then becomes single-valued. At the “cut”, left of the point $-1/2$, the two roots ζ_1 and ζ_2 are interchanged, so we must imagine that the upper sheet for ζ_1 continues into the lower sheet for ζ_2 (shaded in Fig. 4.1) and vice-versa. If we denote the manifold obtained in this way by M , then the map

$$\begin{cases} M \longrightarrow \mathbb{C} \\ \mu \longmapsto \zeta \end{cases} \quad (4.3)$$

becomes an everywhere continuous and holomorphic map (with the exception of the pole). M is then called the *Riemann surface* of the algebraic function $\mu \mapsto \zeta$.

A three-dimensional view of the map

$$\begin{cases} M \longrightarrow \mathbb{R} \\ \mu \longmapsto |\zeta| \end{cases} \quad (4.4)$$

is represented in Fig. 4.2.

More General Methods. Most methods of Sect. V.3 are so-called *multistep Runge-Kutta methods* defined by the formulas

$$y_{n+k} = \sum_{j=1}^k a_j y_{n+j-1} + h \sum_{j=1}^s b_j f(x_n + c_j h, v_j^{(n)}) \quad (4.5a)$$

$$v_i^{(n)} = \sum_{j=1}^k \tilde{a}_{ij} y_{n+j-1} + h \sum_{j=1}^s \tilde{b}_{ij} f(x_n + c_j h, v_j^{(n)}). \quad (4.5b)$$

This is *the* subclass of general linear methods (Example III.8.5) for which the external stages represent the solution $y(x)$ on an equidistant grid. The bulk of numerical work for applying the above method are the implicit stages (4.5b).

For the stability analysis we set as now usual $f(x, y) = \lambda y$, $h\lambda = \mu$ and $(y_n, y_{n+1}, \dots, y_{n+k}) = (1, \zeta, \dots, \zeta^k)$. Equation (4.5b) then becomes in vector notation (using $\vec{\zeta} = (1, \zeta, \dots, \zeta^{k-1})^T$)

$$\vec{v} = (I - \mu \tilde{B})^{-1} \tilde{A} \vec{\zeta}, \quad (4.6)$$

which is rational in μ with denominator $\det(I - \mu\tilde{B})$. Inserting this into (4.5a) and multiplying with this denominator we obtain a characteristic equation of the form

$$Q(\mu, \zeta) \equiv q_k(\mu)\zeta^k + q_{k-1}(\mu)\zeta^{k-1} + \dots + q_0(\mu) = 0 \quad (4.7)$$

where $q_k(\mu) = \det(I - \mu\tilde{B})$ and all $q_j(\mu)$ are polynomials in μ of degree $\leq s$.

Multiderivative multistep methods, on the other hand, may be written as (M. Reimer 1967, R. Jeltsch 1976)

$$\sum_{j=0}^s h^j \sum_{i=0}^k \alpha_{ij} D^j y_{n+i} = 0 \quad (4.8)$$

where the computation of higher derivatives $D^j y$ is done by Eq. (II.12.3). For the equation $y' = \lambda y$ we have $D^j y = \lambda^j y$ and inserting this into (4.8) together with $(y_n, y_{n+1}, \dots, y_{n+k}) = (1, \zeta, \dots, \zeta^k)$ we obtain at once a characteristic equation of the form (4.7). Here, the degree s of the polynomials $\varphi_j(\mu)$ is equal to the order of the highest derivative taken. The bulk of numerical work for evaluating (4.8) is the determination of y_{n+k} from an implicit equation containing y_{n+k} , $Dy_{n+k}, \dots, D^s y_{n+k}$. If the last of these derivatives is present (i.e., if $\alpha_{ks} \neq 0$), then the degree of $q_k(\mu)$ in (4.7) will be s .

The Riemann surface M of (4.7) will consist of k sheets, one for each of the k roots ζ_j . The *branch points* are values of μ for which two or several roots of (4.7) coalesce to an m -fold root. These are the roots of a certain "discriminant" (see any classical book on Algebra, e.g., the famous "Weber", Vol. I, § 50); hence for irreducible $Q(\mu, \zeta)$ there are only a finite number of such points. The movement of the coalescing roots ζ_j , when μ surrounds such a branch point, has been carefully studied by Puiseux: They usually form what Puiseux calls a "système circulaire", i.e., they are cyclically permuted at each revolution like the values of the complex function $\sqrt[m]{z}$ near the origin. The Riemann surface must then follow these "monodromies" and must be cut along certain lines and rejoined appropriately. The location of these cuts is not unique.

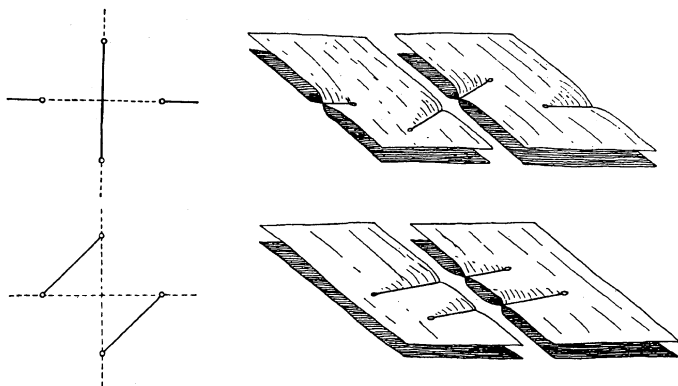


Fig. 4.3. Different cuts for (4.9) (Hurwitz & Courant 1925)

Different possibilities for cutting the Riemann surface of, say, the function

$$\zeta^2 - (1 - \mu^4) = 0 \quad (4.9)$$

with branch points at ± 1 and $\pm i$, are shown in a classical figure reproduced from the book of Hurwitz & Courant, second edition 1925, p. 360 (Fig. 4.3).

Poles Representing Numerical Work

Only 85 miles (geog.) from the Pole, but it's going to be a stiff pull *both ways* apparently; still we do make progress, which is something. (R.F. Scott, January 10, 1912; first mention of interrelation between poles and stiffness in the literature)

We have just seen that the degree s of $q_k(\mu)$ in (4.7) expresses the numerical work (either the number of implicit stages or the number of derivatives for the implicit solution). Now $q_k(\mu)$ will possess s zeros $\mu_1, \mu_2, \dots, \mu_s$. What happens if μ approaches one of these zeros? The polynomial (4.7) of degree k (with k roots $\zeta_1(\mu), \dots, \zeta_k(\mu)$) suddenly becomes a polynomial of degree $k-1$ with only $k-1$ roots. Where does the last one go? Well, by Vieta's Theorem, it must go to infinity. In order to compute its asymptotic behaviour, suppose $q_k(\mu_0) = 0$, $q'_k(\mu_0) \neq 0$, $q_{k-1}(\mu_0) \neq 0$ and that ζ is large. Then all terms $q_{k-2}(\mu)\zeta^{k-2}, \dots, q_0(\mu)$ are dominated by $q_{k-1}(\mu)\zeta^{k-1}$ and may be neglected. It results that

$$\zeta \sim -\frac{q_{k-1}(\mu_0)}{q'_k(\mu_0)} \frac{1}{\mu - \mu_0} \quad \text{as } \mu \rightarrow \mu_0, \quad (4.10)$$

hence the algebraic function $\zeta(\mu)$ possesses a pole on *one* of its sheets. If $q_k(\mu_0) = 0$ is a multiple root, the corresponding pole will be multiple too.

It is also possible that the pole in question coincides with a branch point. This happens when in addition to $q_k(\mu_0) = 0$ also $q_{k-1}(\mu_0) = 0$. In this case *two* roots $\zeta_j(\mu)$ tend to infinity, but *more slowly*, like $\pm C(\mu - \mu_0)^{-1/2}$ (Exercise 1). We therefore count both “half-poles” together as *one* pole again. If c is a boundary curve of a neighbourhood V of μ_0 (which around this branch point surrounds μ_0 twice before closing up), the argument of $\zeta(\mu)$ makes just *one* clockwise revolution on this path. Fig. 4.4 illustrates this fact with an example.

Recapitulating we may state:

Lemma 4.1. *The Riemann surface for the characteristic equation of a multistep Runge-Kutta method with s implicit stages per step (or a multiderivative multistep method with s implicit derivative evaluations) includes at most s poles of the algebraic function $\zeta(\mu)$. \square*

We shall see below that Lemma 4.1 remains true for the whole class of general linear methods, but for the moment we are “impatient et joyeux d’aller au combat”

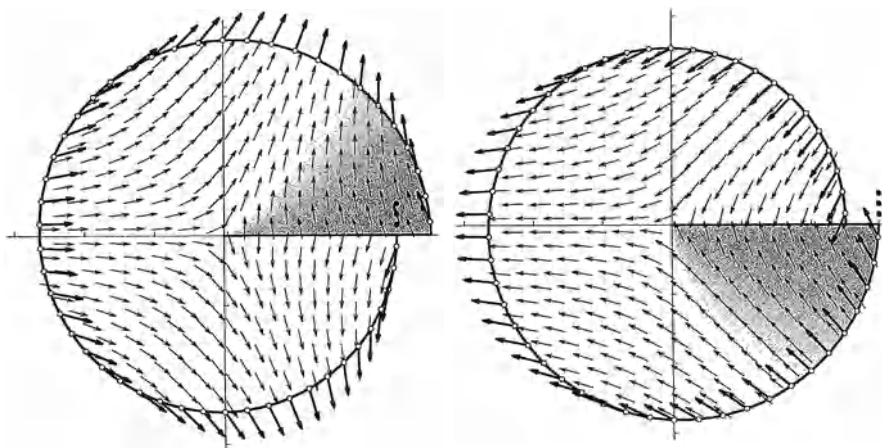


Fig. 4.4. Behaviour of roots of $\mu\zeta^2 + 2\mu\zeta + 2 - \mu = 0$ near the origin $\mu = 0$

(Astérix Légionnaire, pp. 29 and 30). The *argument principle* also remains valid on Riemann surfaces and we state it as follows:

“On the left, isn’t it ?” — “Right.”
 “On the right ?” — “Left, leeeft!”
 (John Cleese in “Clockwise”)

Lemma 4.2. *Suppose that a domain $F \subset M$ contains no zeros of $\zeta(\mu)$ and that its boundary consists of closed loops $\gamma_1, \dots, \gamma_\ell$. Then the number of poles of $\zeta(\mu)$ contained in F is equal to the total number of clockwise revolutions of $\arg(\zeta(\mu))$ along $\gamma_1, \dots, \gamma_\ell$, each passed through in that direction which leaves F to the left of γ_j .*

The *proof* is by cutting F into thousand pieces, each of which is homeomorphic to a disc in \mathbb{C} , and by adding up all revolution numbers which cancel along the cuts, because the adjacent edges are traversed in opposite directions. \square

Order and Order Stars

... denn das Klare und leicht Faßliche zieht uns an, das Verwickelte schreckt uns ab. (D. Hilbert, Paris 1900)

Guided by the ideas of Sect. IV.4, we now compare the absolute values of the characteristic roots $|\zeta_1|$ and $|\zeta_2|$ for the BDF2 scheme (4.2) with the exponential function $|e^\mu| = e^{\operatorname{Re} \mu}$, hence we define (Wanner, Hairer & Nørsett 1978)

$$A_j = \left\{ \mu \in \mathbb{C} ; |\zeta_j(\mu)| > |e^\mu| \right\} \quad j = 1, 2. \quad (4.11)$$

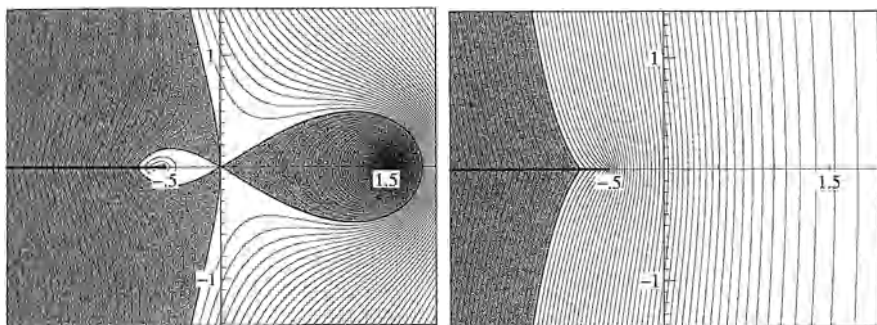


Fig. 4.5. The order star (4.14) for BDF2

These sets, on precisely the same scale as in Fig. 4.1, are represented in Fig. 4.5.

The sets A_j continue across the cuts in the same way as do the roots, it is therefore natural to embed them into the Riemann surface M and define

$$A = \left\{ \mu \in M ; |\zeta(\mu)| > |e^{\pi(\mu)}| \right\} \quad (4.12)$$

where $\pi : M \rightarrow \mathbb{C}$ is the natural projection.

Fig. 4.5 shows clearly an order star with three sectors for $\zeta_1(\mu)$, but none for $\zeta_2(\mu)$, and we guess that this has to do with the order of the method, which is two. Lemma 4.3 below will extend Lemma IV.4.3 to multistep methods.

By putting $h = 0$ in (4.5) (hence $\mu = 0$ in (4.7)), and

$$(y_n, y_{n+1}, \dots, y_{n+k-1}) = (1, 1, \dots, 1)$$

(hence $\zeta = 1$ in (4.7)), we must have by consistency that $y_{n+k} = 1$ too, i.e., that $Q(0, 1) = 0$. This corresponds to the formula $\varrho(1) = 0$ in the multistep case (see (III.2.6)). But for $h = 0$ the difference equation (4.5a) is stable only if $\zeta = 1$ is a *simple* root of the polynomial equation $Q(0, \zeta) = 0$. Hence we must have

$$Q(0, 1) = 0, \quad \frac{\partial Q}{\partial \zeta}(0, 1) \neq 0. \quad (4.13)$$

The analytic continuation $\zeta_1(\mu)$ of this root in the neighbourhood of the origin (as far as it is not embarrassed with branch points) will be called the *principal root*, the corresponding surface the *principal sheet* of M .

Lemma 4.3. *For stable multistep Runge-Kutta (or multiderivative) methods of order p the set A possesses a star of $p+1$ sectors on the principal sheet in the neighbourhood of the origin.*

Proof. We fix $\lambda \in \mathbb{C}$, set $y' = \lambda y$ and take for y_0, \dots, y_{k-1} exact initial values $1, e^\mu, \dots, e^{(k-1)\mu}$. The order of the method then tells us that the *local error* (see Fig. III.2.1), i.e., the difference between $e^{k\mu}$ and the numerical solution y_k computed from (4.5a), must be $\tilde{C} \cdot h^{p+1}$ for $h \rightarrow 0$, hence $\tilde{C} \lambda^{-p-1} \mu^{p+1}$ for $\mu \rightarrow 0$.

Thus, Formula (4.5) with *all* y_j replaced by $e^{j\mu}$ will lead to

$$Q(\mu, e^\mu) = \overline{C}\mu^{p+1} + \mathcal{O}(\mu^{p+2}). \quad (4.14)$$

We subtract (4.14) from (4.7), choose for $\zeta(\mu)$ the principal root $\zeta_1(\mu)$ (for which $e^\mu - \zeta_1(\mu)$ is small for $|\mu|$ small) and linearize. This gives

$$\frac{\partial Q}{\partial \zeta}(0, 1) \cdot (e^\mu - \zeta_1(\mu)) = \overline{C}\mu^{p+1} + \dots$$

and by dividing through by the non-zero constant (4.13)

$$e^\mu - \zeta_1(\mu) = C \cdot \mu^{p+1} + \mathcal{O}(\mu^{p+2}) \quad \text{for} \quad \mu \rightarrow 0. \quad (4.15)$$

The rest of the proof now goes exactly analogously to that of Lemma IV.4.3. There is also not much difference in the case of multiderivative methods. \square

The constant C of (4.15) is called the *error constant* of the method. This is consistent with Formula (III.2.6) and (III.2.13) for multistep methods and with (IV.3.5) for Runge-Kutta methods.

The stability domain of multistep Runge-Kutta methods as well as their A -stability is defined in the same way as for multistep methods (see Definition 1.1). One has only to interpret $\zeta_1(\mu), \dots, \zeta_k(\mu)$ as the roots of (4.7).

The “Daniel and Moore Conjecture”

It is conjectured here that no A -stable method of the form of Eq. 5-6 can be of order greater than $2J+2$ and that, of those A -stable methods of order $2J+2$, the smallest error constant is exhibited by the *Hermite method* . . .

(Daniel & Moore 1970, p. 80)

At the time when no simple proof for Dahlquist’s second barrier was known, a proof of its generalization, the Daniel & Moore conjecture, seemed quite hopeless. Y. Genin (1974) constructed A -stable multistep multiderivative methods with astonishingly high “order” contradicting the conjecture. R. Jeltsch (1976) later cleared up the mystery by showing that Genin’s methods had 1 as multiple root of $\varrho(\zeta)$ and hence the “effective” order was lower. The conjecture was finally proved in 1978 with the help of order stars:

Theorem 4.4. *The highest order of an A -stable s -stage Runge-Kutta (or s -derivative) multistep method is $2s$. For the A -stable methods of order $2s$ the error constant satisfies*

$$(-1)^s C \geq \frac{s! s!}{(2s)! (2s+1)!}. \quad (4.16)$$

Proof. By A -stability, we have for *all* roots $|\zeta_j(iy)| \leq 1$ along the imaginary axis; hence the order star A is nowhere allowed to cross the imaginary axis. We consider $A^+ = A \cap \pi^{-1}(\mathbb{C}^+)$, the part of the order star which lies above \mathbb{C}^+ . As in Lemma IV.4.4, A^+ must be finite on *all* sheets of M . The boundary of A^+ may consist of several closed curves. As in Lemma IV.4.5, the argument of $\zeta(\mu)/e^\mu$ is steadily increasing along ∂A^+ . Since at the origin we have a star with $p+1$ sectors (Lemma 4.3), of which at least $\lceil \frac{p+1}{2} \rceil$ lie in \mathbb{C}^+ , the boundary curves of A^+ must visit the origin at least $\lceil \frac{p+1}{2} \rceil$ times. Hence the total rotation number is at least $\lceil \frac{p+1}{2} \rceil$ and from Lemmas 4.1 and 4.2 we conclude that

$$\left\lceil \frac{p+1}{2} \right\rceil \leq s. \quad (4.17)$$

This implies that $p \leq 2s$ and the first assertion is proved.

We now need a new idea for the part concerning the error constant. The following reasoning will help: the star A expresses the fact that the surface $|\zeta(\mu)/e^\mu|$ goes up and down around the origin like Montaigne's ruff. There, the *error constant* has to do with the *height* of these waves. So if we want to compare different error constants we must compare $|\zeta(\mu)/e^\mu|$ to $|R(\mu)/e^\mu|$, where $R(\mu)$ is the characteristic function of a second method. By dividing the two expressions, e^μ cancels and we define



$$B = \left\{ \mu \in M ; \left| \frac{\zeta(\mu)}{R(\pi(\mu))} \right| > 1 \right\}, \quad (4.18)$$

called the *relative order star*. For $R(z)$ we choose the diagonal Padé approximation $R_{ss}(z)$ with s zeros and s poles (see (IV.3.30)). By subtracting (IV.3.31) (with $j = k = s$) from (4.15) (where it is now supposed that $p = 2s$) we obtain

$$R_{ss}(\mu) - \zeta_1(\mu) = \underbrace{\left(C - (-1)^s \frac{s! s!}{(2s)!(2s+1)!} \right)}_{\tilde{C}} \mu^{2s+1} + \dots \quad (4.19)$$

It is known that $|R_{ss}(iy)| = 1$ for all $y \in \mathbb{R}$ and that all zeros of $R_{ss}(z)$ lie in \mathbb{C}^- (Theorem IV.4.12). Therefore the set B in (4.18) cannot cross the imaginary axis (as before) and the quotient $|\zeta(\mu)/R(\pi(\mu))|$ has no other poles above \mathbb{C}^+ than those of $\zeta(\mu)$, of which, we know, there are at most s . Therefore the sectors of the relative order star B must exhibit the same colours as those of the classical order star A for diagonal Padé (see Fig. IV.4.2). Otherwise an extra pole would be needed. We conclude that the error constants must have the same sign (see Lemma IV.4.3), hence (see IV.3.31) $(-1)^s \tilde{C} > 0$, which leads to (4.16).

Equality $\tilde{C} = 0$ would produce an order star B of even higher order which is impossible with s poles, unless the two methods are identical. \square

Remarks. a) The first half is in fact superfluous, since the inequality (4.16) implies that the $2s$ -th order error constant $C \neq 0$, hence necessarily $p \leq 2s$. It has been retained for its beauty and simplicity, and for readers who do not want to study the second half.

b) The proof never uses the full hypothesis of A -stability; the only property used is stability on the imaginary axis (I -stability, see (IV.3.6)). Thus Theorem 4.4 allows the following sharpening, which then extends Theorem IV.4.7 to multistep methods:

Theorem 4.5. *Suppose that an I -stable s -stage Runge-Kutta (or s -derivative) multistep method possesses a characteristic function $\zeta(\mu)$ with s_1 poles in \mathbb{C}^+ . Then*

$$p \leq 2s_1 \quad (4.20)$$

and the error constant for all such I -stable methods of order $p = 2s_1$ satisfies

$$(-1)^{s_1} C \geq \frac{s_1! s_1!}{(2s_1)! (2s_1 + 1)!}. \quad (4.21)$$

□

Another interpretation of this theorem is the following result (compare with Theorem IV.4.8), which in the case $s = 1$ is due to R. Jeltsch (1978).

Theorem 4.6. *Suppose that an I -stable method with s poles satisfies $p \geq 2s - 1$. Then it is A -stable.*

Proof. If only $s - 1$ poles were in \mathbb{C}^+ , we would have $p \leq 2s - 2$, a contradiction. Hence all poles of $\zeta(\mu)$ are in \mathbb{C}^+ and A -stability follows from the maximum principle. □

Methods with Property C

It is now tempting to extend the proof of Theorem 4.4 to *any* method other than the diagonal Padé method. But this meets with an essential difficulty in defining (4.18) if $R(\mu)$ is a multistep method defined on *another* Riemann surface, since then the definition of B makes no sense. The following observation will help: The second part of the proof of Theorem 4.4 only took place in \mathbb{C}^+ , which was the *instability* domain of the “comparison method”. This leads to

Definition 4.7 (Jeltsch & Nevanlinna 1982). Let a method be given with characteristic polynomial (4.7) satisfying (4.13) and denote its stability domain by S_R . We say that this method has *Property C* if the principal sheet includes no branch points outside of $\pi^{-1}(S_R)$ (with ∞ included if S_R is bounded), and the principal root $R_1(\mu)$ produces the whole instability of the method, i.e.,

$$\Delta_R := \partial S_R = \{\mu \in \mathbb{C} ; |R_1(\mu)| = 1\}. \quad (4.22)$$

Examples. All one-step methods have Property C , of course. Linear multistep methods whose root locus curve is simply closed have Property C too. In this situation all roots except $R_1(\mu)$ have modulus smaller than one for all $\mu \notin \pi^{-1}(S_R)$. Thus the principal sheet cannot have a branch point there. The explicit 4th order Adams method analyzed in Fig. 1.1 does *not* have Property C . The implicit Adams methods (see Fig. 1.3) have Property C for $k \leq 5$. Also, the 4th order implicit Milne-Simpson method (1.16) has property C .

Definition 4.7 allows us to replace $R_{ss}(\mu)$ in the proof of Theorem 4.4 by $R_1(\mu)$, \mathbb{C}^+ by the exterior of S_R , the imaginary axis by Δ_R and to obtain the following theorem (Jeltsch and Nevanlinna the 5th of April, 1979 at 5 a.m. in Champaign; G.W. the 5th of April, 1979 at 4.30 a.m. in Urbana. How was this coincidence possible? E-mail was not yet in general use at that time; was it Psi-mail?)

Theorem 4.8. *Let a method with characteristic function $R(\mu)$, stability domain S_R and order p_R possess Property C . If another method with characteristic function $\zeta(\mu)$, stability domain S_ζ and order p_ζ is more stable than R , i.e., if*

$$S_\zeta \supset S_R, \quad (4.23)$$

then

$$p \leq 2s \quad (4.24)$$

where

$$p = \min(p_R, p_\zeta) \quad (4.25)$$

and s is the number of poles of $\zeta(\mu)$, each counted with its multiplicity, which are not poles of the principal root $R_1(\mu)$ of $R(\mu)$. \square

... and tried to optimize the stability boundary. Despite many efforts we were not able to exceed $\sqrt{3}$, the stability boundary of the Milne-Simpson method ... (K. Dekker 1981)

As an illustration of Theorem 4.8 we ask for the largest stability interval on the imaginary axis $I_r = [-ir, ir] \subset \mathbb{C}$ of a 3rd order multistep method (for hyperbolic equations). Since we have $s = 1$ for linear multistep methods, $p = 3$ contradicts (4.24) and we obtain from Theorem 4.8 by using for $R(\mu)$ the Milne-Simpson method (1.16):

Theorem 4.9 (Dekker 1981, Jeltsch & Nevanlinna 1982). *If a linear multistep method of order $p \geq 3$ is stable on I_r , then $r \leq \sqrt{3}$.* \square

The second part of Theorem 4.4 also allows an extension, the essential ingredient for its proof has been the sign of the error constant for the diagonal Padé approximation.

Theorem 4.10. Consider a method with characteristic equation (4.7) satisfying (4.13) and let p denote its order and C its error constant. Suppose

- a) the method possesses Property C ,
- b) the principal root $R_1(\mu)$ possesses s poles,
- c) $\text{sign}(C) = (-1)^s$
- d) $p \geq 2s - 1$.

Then this method is “optimal” in the sense that every other method with s poles which is stable on Δ_R of (4.22) has either lower order or, for the same order, a larger (in absolute value) error constant. \square

Examples. The diagonal and first sub-diagonal Padé approximations satisfy the above hypotheses (see Eq. (IV.3.30)). Also I -stable linear multistep methods with Property C can be applied.

Remark 4.11. Property C allows the extension of Theorem IV.4.17 of Jeltsch & Nevanlinna to explicit multistep methods. Thus explicit methods with comparable numerical work cannot have including stability domains. Exercise 4 below shows that Property C is a necessary condition. Remember that explicit methods have all their poles at infinity.

General Linear Methods

The large class of general linear methods (Example III.8.5) written in obvious matrix notation

$$v_n = \tilde{A}u_n + h\tilde{B}f(v_n) \quad (4.26a)$$

$$u_{n+1} = Au_n + hBf(v_n) \quad (4.26b)$$

seems to allow much more freedom to break the Daniel & Moore conjecture. This is not the case as we shall see in the sequel.

The bulk of numerical work for solving (4.26) is represented by the implicit stages (4.26a) and hence depends on the structure of the matrix \tilde{B} . Inserting $y' = \lambda y$ leads to

$$u_{n+1} = S(\mu)u_n \quad (4.27)$$

where

$$S(\mu) = A + \mu B(I - \mu\tilde{B})^{-1}\tilde{A}. \quad (4.28)$$

The stability of the numerical method (4.27) is thus governed by the eigenvalues of the matrix $S(\mu)$. The elements of this matrix are seen to be rational functions in μ .

Lemma 4.12. *If the characteristic polynomial of $S(\mu)$ is multiplied by $\det(I - \mu\tilde{B})$ then it becomes polynomial in μ :*

$$\begin{aligned} \det(\zeta I - S(\mu)) \cdot \det(I - \mu\tilde{B}) &= q_k(\mu)\zeta^k + q_{k-1}(\mu)\zeta^{k-1} + \dots + q_0(\mu) \\ &=: Q(\mu, \zeta) \end{aligned} \quad (4.29)$$

where q_0, \dots, q_k are polynomials of degree $\leq s$ and $q_k(\mu) = \det(I - \mu\tilde{B})$.

Proof. Suppose first that \tilde{B} is diagonalizable as

$$T^{-1}\tilde{B}T = \text{diag}(\beta_1, \dots, \beta_s) \quad (4.30)$$

so that from (4.28)

$$S(\mu) = A + B T \text{diag}(w_1, \dots, w_s) T^{-1} \tilde{A} = A + \sum_{i=1}^s w_i \vec{d}_i \vec{c}_i^T \quad (4.31)$$

where

$$\left. \begin{aligned} w_i &= \frac{\mu}{1 - \mu\beta_i} \\ \vec{d}_i &= i\text{-th column of } BT \\ \vec{c}_i^T &= i\text{-th row of } T^{-1}\tilde{A}. \end{aligned} \right\} \quad i = 1, \dots, s \quad (4.32)$$

We write the matrix $\zeta I - S(\mu)$ in terms of its column vectors

$$\left(\zeta \vec{e}_1 - \vec{a}_1 - w_1 c_{11} \vec{d}_1 - w_2 c_{12} \vec{d}_2 - \dots, \zeta \vec{e}_2 - \vec{a}_2 - w_1 c_{21} \vec{d}_1 - w_2 c_{22} \vec{d}_2 - \dots, \dots \right).$$

Its determinant, the characteristic polynomial of $S(\mu)$, is computed using the multilinearity of \det and considering ζ, w_i, c_{ij} as scalars. All terms containing one of the w_j to any power higher than 1 cancel, because the corresponding factor is a determinant with two or more identical columns. Thus, if $\det(\zeta I - S(\mu))$ is multiplied by $\prod_{i=1}^s (1 - \mu\beta_i) = \det(I - \mu\tilde{B})$ it becomes a polynomial of the form (4.29).

A non-diagonalizable matrix \tilde{B} is considered as the limit of diagonalizable matrices. The coefficients of the polynomial $Q(\mu, \zeta)$ depend continuously on \tilde{B} . \square

We conclude that Lemma 4.1 again remains valid for general linear methods. The s poles on the Riemann surface for the algebraic function $Q(\mu, \zeta) = 0$ are located at the positions $\mu = 1/\beta_1, \dots, \mu = 1/\beta_s$ where β_i are the eigenvalues of the matrix \tilde{B} .

We next have to investigate the *order conditions*, i.e., the analogue of Lemma 4.3. Recall that general linear methods must be equipped with a *starting procedure* (see Eq. (III.8.4a)) which for the differential equation $y' = \lambda y$ will be of the form $u_0 = \psi(\mu) \cdot y_0$ with $\psi(0) \neq 0$. Here $\mu = h\lambda$ and $\psi(\mu)$ is a k -vector of

polynomials or rational functions of μ . Then the diagram of Fig. III.8.1 becomes the one sketched in Fig. 4.6.

The order condition (see Formula (III.8.16) of Lemma III.8.11) then gives:

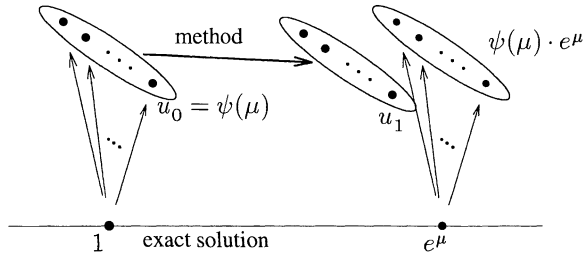


Fig. 4.6. General linear method for $y' = \lambda y$

Lemma 4.13. *If the general linear method (4.26) is of order p then*

$$(e^\mu I - S(\mu))\psi(\mu) = \mathcal{O}(\mu^p) \quad \text{for } \mu \rightarrow 0 \quad (4.33a)$$

$$E(e^\mu I - S(\mu))\psi(\mu) = \mathcal{O}(\mu^{p+1}) \quad \text{for } \mu \rightarrow 0 \quad (4.33b)$$

where E is defined in (III.8.12) and $S(\mu)$ is given in (4.28). \square

Formula (4.33) tells us, roughly, that $\psi(\mu)$ is an approximate eigenvector of $S(\mu)$ with eigenvalue e^μ . We shall now see how this information can be turned into order conditions for the correct eigenvalues of $S(\mu)$.

Definition 4.14. Let ℓ be the number of principal sheets of (4.29), i.e., the multiplicity of 1 as eigenvalue of $S(0)$ (which, by stability, must then be a simple root of the minimal polynomial). ℓ is also the dimension of I in (III.8.12) and the rank of E .

Theorem 4.15. *Suppose that there exists $\psi(\mu)$ with $\psi(0) \neq 0$ such that the general linear method satisfies the conditions (4.33) for order $p \geq 1$. Then the ℓ -fold eigenvalue 1 of S continues into ℓ eigenvalues $\zeta_j(\mu)$ of $S(\mu)$ which satisfy*

$$e^\mu - \zeta_j(\mu) = \mathcal{O}(\mu^{p_j+1}) \quad \mu \rightarrow 0 \quad (4.34)$$

with

$$p_j \geq 0, \quad \sum_{j=1}^{\ell} p_j \geq p. \quad (4.35)$$

Examples. a) The matrix

$$S(\mu) = \begin{pmatrix} 1 + \mu & \frac{20}{9}\mu^2 \\ 3\mu + \frac{11}{80}\mu^2 & 1 - \frac{37}{3}\mu + \frac{13}{3}\mu^2 \end{pmatrix} \quad (4.36)$$

has $\ell = 2$ so that $E = I$ in (4.33b). There is a vector $\psi(\mu)$ (non-vanishing for $\mu = 0$) such that

$$(e^\mu I - S(\mu))\psi(\mu) = \mathcal{O}(\mu^6),$$

i.e., $p = 5$. The eigenvalues

$$\zeta_{1,2}(\mu) = \left(1 - \frac{17\mu}{3} + \frac{13\mu^2}{6}\right) \pm \frac{20\mu}{3} \sqrt{1 - \frac{\mu}{2} + \frac{9\mu^2}{80}}$$

satisfy $e^\mu - \zeta_1(\mu) = \mathcal{O}(\mu^6)$, $e^\mu - \zeta_2(\mu) = \mathcal{O}(\mu)$, which is (4.34) with $p_1 = 5$, $p_2 = 0$.

b) The matrix

$$S(\mu) = \begin{pmatrix} 1 + 2\mu + \frac{\mu^2}{2} & -\mu \\ \mu & 1 + \frac{\mu^2}{2} \end{pmatrix} \quad (4.37)$$

satisfies (4.33) with $\ell = 2$, $p = 4$. Its eigenvalues $\zeta_{1,2}(\mu) = 1 + \mu + \mu^2/2$ fulfil (4.34) with $p_1 = p_2 = 2$.

c) The example

$$S(\mu) = \begin{pmatrix} 1 + 2\mu & -\mu + \mu^2 \\ \mu & 1 \end{pmatrix} \quad (4.38)$$

has $\ell = 2$, $p = 1$ in (4.33). Its eigenvalues $\zeta_{1,2}(\mu) = 1 + \mu \pm \sqrt{\mu^3}$ satisfy (4.34) with $p_1 = p_2 = 1/2$. This example shows that the p_j in (4.34) need not be integers.

Proof of Theorem 4.15. We introduce the matrix

$$\tilde{S}(\mu) = e^\mu I - S(\mu) \quad (4.39)$$

which has the same eigenvectors as $S(\mu)$ and the corresponding eigenvalues

$$\tilde{\zeta}_j(\mu) = e^\mu - \zeta_j(\mu). \quad (4.40)$$

Formulas (4.34) and (4.35) now say simply that

$$\prod_{j=1}^{\ell} \tilde{\zeta}_j(\mu) = \mathcal{O}(\mu^{p+\ell}) \quad \mu \rightarrow 0. \quad (4.41)$$

Since the product of the eigenvalues is, as we know, the determinant of the matrix, we look for information about $\det \tilde{S}(\mu)$.

After a suitable change of coordinates (via the transformation matrix T of (III.8.12)) we suppose the matrix $S = S(0)$ in Jordan canonical form. We then separate blocks of dimensions ℓ and $k - \ell$ so that

$$E = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \quad S(\mu) = \begin{pmatrix} I + \mathcal{O}(\mu) & \mathcal{O}(\mu) \\ \mathcal{O}(\mu) & \mathcal{O}(1) \end{pmatrix}, \quad \psi(\mu) = \begin{pmatrix} \psi_1(\mu) \\ \psi_2(\mu) \end{pmatrix} \quad (4.42)$$

$$\tilde{S}(\mu) = \begin{pmatrix} \tilde{S}_{11}(\mu) & \tilde{S}_{12}(\mu) \\ \tilde{S}_{21}(\mu) & \tilde{S}_{22}(\mu) \end{pmatrix} = \begin{pmatrix} \mathcal{O}(\mu) & \mathcal{O}(\mu) \\ \mathcal{O}(\mu) & \mathcal{O}(1) \end{pmatrix} \quad (4.43)$$

where it is important to notice that $\tilde{S}_{22}(0)$ is invertible; this is because E collects all eigenvalues equal to 1, thus $S_{22}(0)$ has no eigenvalues equal to 1 and $\tilde{S}_{22}(0)$ has none equal to zero. Conditions (4.33) now read

$$\begin{pmatrix} \tilde{S}_{11}(\mu) & \tilde{S}_{12}(\mu) \\ \tilde{S}_{21}(\mu) & \tilde{S}_{22}(\mu) \end{pmatrix} \begin{pmatrix} \psi_1(\mu) \\ \psi_2(\mu) \end{pmatrix} = \begin{pmatrix} \mathcal{O}(\mu^{p+1}) \\ \mathcal{O}(\mu^p) \end{pmatrix}. \quad (4.44)$$

Putting $\mu = 0$ in (4.44) we get $\psi_2(0) = 0$. The assumption $\psi(0) \neq 0$ thus implies that at least one component of $\psi_1(0)$, say the j -th component $\psi_{1j}(0)$, does not vanish. Cramer's rule then yields

$$\det \tilde{S}(\mu) \cdot \psi_{1j}(\mu) = \det T(\mu), \quad (4.45)$$

where $T(\mu)$ is obtained from $\tilde{S}(\mu)$ by replacing its j -th column by the right-hand side of (4.44). One easily sees that $\det T(\mu) = \mathcal{O}(\mu^{p+\ell})$ (take out a factor μ from each of the first ℓ lines and a factor μ^p from the j -th column). Because of $\psi_{1j}(0) \neq 0$ this implies $\det \tilde{S}(\mu) = \mathcal{O}(\mu^{p+\ell})$. We have thus proved (4.41) (hence (4.34) and (4.35)), because $\tilde{\zeta}_{\ell+1}, \dots, \tilde{\zeta}_k$ do not converge to zero for $\mu \rightarrow 0$. \square

The next lemma excludes fractional orders for A -stable methods:

Lemma 4.16. *For I -stable general linear methods the orders p_j in (4.34) must be integers.*

Proof. Divide (4.34) by e^μ , let

$$\frac{\zeta_j(\mu)}{e^\mu} = 1 - C\mu^{m/r} + \dots \quad (4.46)$$

where $p_j + 1 = m/r$, and suppose that $r > 1$ and m, r are relatively prime. Since $e^\mu - \zeta_j(\mu)$ are the eigenvalues of the matrix (4.39), hence the roots of an analytic equation, the presence of a root $\mu^{m/r}$ involves the occurrence of all branches $\mu^{m/r} \cdot e^{2i\pi j/r}$ ($j = 0, 1, \dots, r-1$). For $\mu = \pm iy = e^{\pm i\pi/2} y$ ($y \in \mathbb{R}$ small), inserted into (4.46), we thus obtain $2r$ different values

$$1 - Cy^{m/r} e^{\pm im\pi/2r} e^{2i\pi j/r} + \dots \quad j = 0, 1, \dots, r-1$$

which form a regular $2r$ -Mercedes star; hence whatever the argument of C is, there are values of $C(\pm iy)^{m/r} e^{2i\pi j/r}$ (for some $0 \leq j \leq r-1$) with negative real part, such that from (4.46) $|\zeta_j(\pm iy)| > 1$. This is a contradiction to I -stability. \square

And here is the “Daniel-Moore conjecture” for general linear methods:

Theorem 4.17. *Let the characteristic function $Q(\mu, \zeta)$ of an I -stable general linear method possess s poles in \mathbb{C}^+ . Then*

$$p \leq 2s. \quad (4.47)$$

Proof. Again we denote by $A^+ = A \cap \pi^{-1}(\mathbb{C}^+)$, the part of the order star lying above \mathbb{C}^+ . By I -stability A^+ does not intersect the imaginary axis $\pi^{-1}(i\mathbb{R})$ on any sheet.

By Theorem 4.15 the boundary curves γ_m of A^+ visit the origin on the different principal sheets at least $\lceil \frac{p_j+1}{2} \rceil$ times ($j = 1, \dots, \ell$) (see (4.17)), where the p_j are integers by Lemma 4.16. Thus by Lemma 4.2

$$\sum_{j=1}^{\ell} \left\lceil \frac{p_j+1}{2} \right\rceil \leq s. \quad (4.48)$$

Multiplying this by 2, using $p_j \leq 2\lceil \frac{p_j+1}{2} \rceil$ and (4.35), we get $p \leq 2s$. \square

Dual Order Stars

Why not interchange the role of the two variables ζ and $\mu \dots$?
(J. Butcher,
June 27, 1989, in West Park Hall, Dundee, at midsummernight)

A -stability implies that for all solutions $\zeta_j(\mu)$ of $Q(\mu, \zeta) = 0$ we have

$$\operatorname{Re} \mu \leq 0 \quad \implies \quad |\zeta_j(\mu)| \leq 1. \quad (4.49)$$

This is logically equivalent to: For all solutions $\mu_j(\zeta)$ of $Q(\mu, \zeta) = 0$ we have

$$|\zeta| \geq 1 \quad \implies \quad \operatorname{Re} \mu_j(\zeta) \geq 0 \quad (4.50)$$

(in fact, pure logic gives us “ $>$ ” on both sides; the “ \geq ” then follow by continuity). Further the order condition (4.15) becomes, by passing to inverse functions for the principal root,

$$\log \zeta - \mu_1(\zeta) = -C(\zeta - 1)^{p+1} + \dots \quad (4.51)$$

Thus order star theory can be very much dualized by the replacements

$$\begin{array}{lll} \text{a)} & \mu & \longleftrightarrow \zeta \\ \text{b)} & 0 & \longleftrightarrow 1 \\ \text{c)} & \text{Imag. axis} & \longleftrightarrow \text{Unit circle} \\ \text{d)} & \operatorname{Re} & \longleftrightarrow |\cdot| \\ \text{e)} & \operatorname{Im} & \longleftrightarrow \operatorname{Arg} \\ \text{f)} & \exp & \longleftrightarrow \log \end{array} \quad (4.52)$$

The analogue of the star defined in (4.12) becomes

$$A = \left\{ \zeta ; \operatorname{Re} \mu(\zeta) \leq \operatorname{Re} (\log \zeta) \right\} = \left\{ \zeta ; \operatorname{Re} \mu(\zeta) \leq \log |\zeta| \right\} \quad (4.53)$$

and the analogue of the relative order star (4.18) becomes

$$B = \left\{ \zeta ; \operatorname{Re} \mu(\zeta) \leq \operatorname{Re} \mu_R(\zeta) \right\}. \quad (4.54)$$

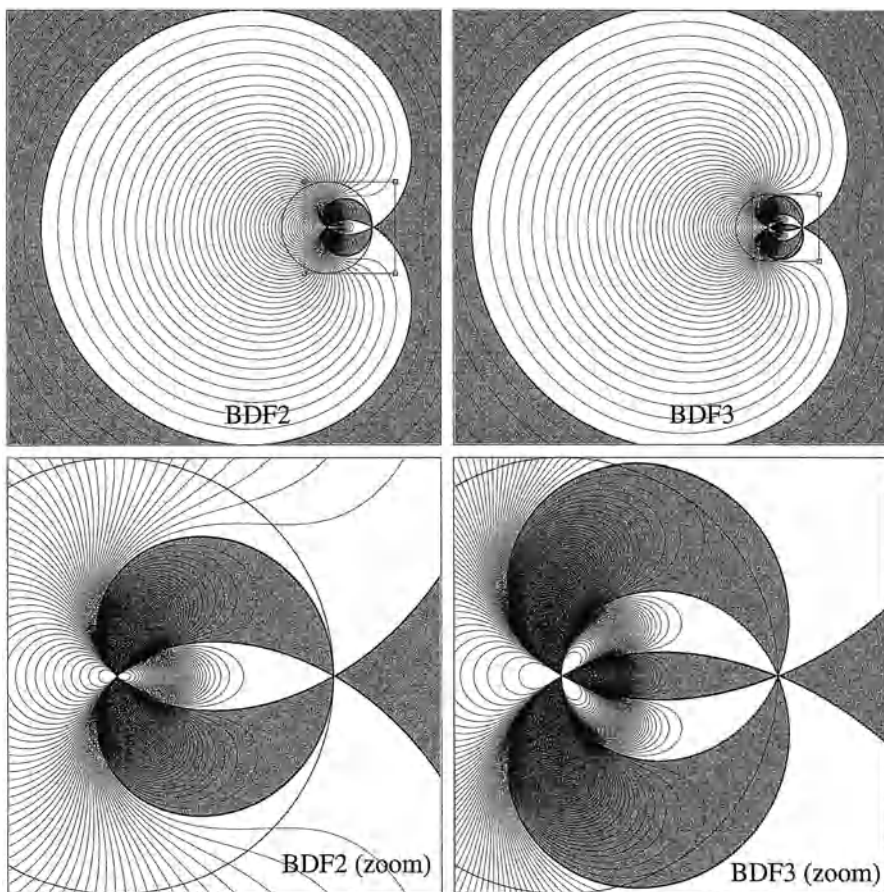


Fig. 4.7. Dual order stars (4.53) for BDF methods

For the special case of the trapezoidal rule this is

$$B = \left\{ \zeta ; \operatorname{Re} \mu(\zeta) \leq \operatorname{Re} \left(2 \frac{\zeta - 1}{\zeta + 1} \right) \right\}. \quad (4.55)$$

The set A is displayed in Fig. 4.7 for the BDF2 and BDF3 methods. It explains once again why A -stable methods of order $> 2s$ are not possible (see Exercise 5).

Still another possibility is to replace (4.50) by the obviously equivalent condition

$$|\zeta| \geq 1 \quad \Rightarrow \quad \operatorname{Re} \frac{1}{\mu_j(\zeta)} \geq 0 \quad (4.56)$$

in which case order condition (4.51) becomes

$$\frac{1}{\log \zeta} - \frac{1}{\mu_1(\zeta)} = C(\zeta - 1)^{p-1} + \dots \quad (4.57)$$

since $\log \zeta$ as well as $\mu_1(\zeta)$ are $(\zeta - 1) + \mathcal{O}((\zeta - 1)^2)$. The order stars now become analogously

$$A = \left\{ \zeta ; \operatorname{Re} \frac{1}{\mu(\zeta)} \geq \operatorname{Re} \frac{1}{\log \zeta} \right\} \quad (4.58)$$

and

$$B = \left\{ \zeta ; \operatorname{Re} \frac{1}{\mu(\zeta)} \geq \operatorname{Re} \frac{1}{\mu_R(\zeta)} \right\}. \quad (4.59)$$

A special advantage of these last definitions is that for linear multistep methods $1/\mu = \sigma(\zeta)/\varrho(\zeta)$, hence the *poles* of the functions involved are the *zeros* of $\varrho(\zeta)$, which play a role in the definition of *ordinary* stability (Sect. III.3). This can be used to obtain a geometric proof of the *first* Dahlquist barrier (Theorem III.3.5), inspired by the paper Iserles & Nørsett (1984) (see Exercise 6).

Also, the proof for Dahlquist's second barrier of Sect. V.1 (Theorem 1.4) can be seen to be nothing else but a study of B of (4.59) where $\mu_R(\zeta)$ represents the trapezoidal rule.

Exercises

1. Analyze the behaviour of the characteristic roots of (4.7) in the neighbourhood of a pole which coincides with a branch point, i.e., solve (4.7) asymptotically for ζ large in the case

$$\varphi_k(\mu_0) = 0, \quad \varphi'_k(\mu_0) \neq 0, \quad \varphi_{k-1}(\mu_0) = 0, \quad \varphi_{k-2}(\mu_0) \neq 0.$$

Show that these roots behave like $\pm C(\mu - \mu_0)^{-1/2}$.

2. Compute the approximate eigenvectors $\psi(\mu)$ such that

$$(e^\mu I - S(\mu))\psi(\mu) = \mathcal{O}(\mu^{p+1})$$

for the matrices $S(\mu)$ given in (4.36), (4.37), (4.38). Show that the stated orders are optimal.

3. Explain with the help of order stars, why the 2-step 2-stage collocation method with $c_2 = 1$ (see Exercise 7 of Sect. V.3) loses A -stability exactly when c_1 crosses the superconvergence point (Exercise 8 of Sect. V.3).
4. Modify the coefficient β in the method

$$y_{n+1} = y_n + h \left(f_n + \frac{1}{2} \nabla f_n + \frac{5}{12} \nabla^2 f_n + \beta \nabla^3 f_n \right),$$

which for $\beta = 3/8$ is the Adams method of order 4, in such a way that the stability domain becomes *strictly* larger. This example shows that the multistep version of Theorem IV.4.17 of Jeltsch & Nevanlinna requires the hypothesis of "Property C".

5. Prove the Daniel & Moore conjecture with the help of the order star A from (4.53).

Hint. The set A is not allowed to cross the unit circle and along the borderlines of A the imaginary part of $\log \zeta - \mu(\zeta)$ must steadily *decrease* (consult (4.52) and the proof of Lemma IV.4.5). Hence a borderline starting and ending at the origin must either pass through a pole (which is not outside the unit circle) or cross the negative real axis in the upward direction (where $\text{Im}(\log \zeta)$ increases by 2π). Since then the set A must be to the left, this is only possible once on each sheet.

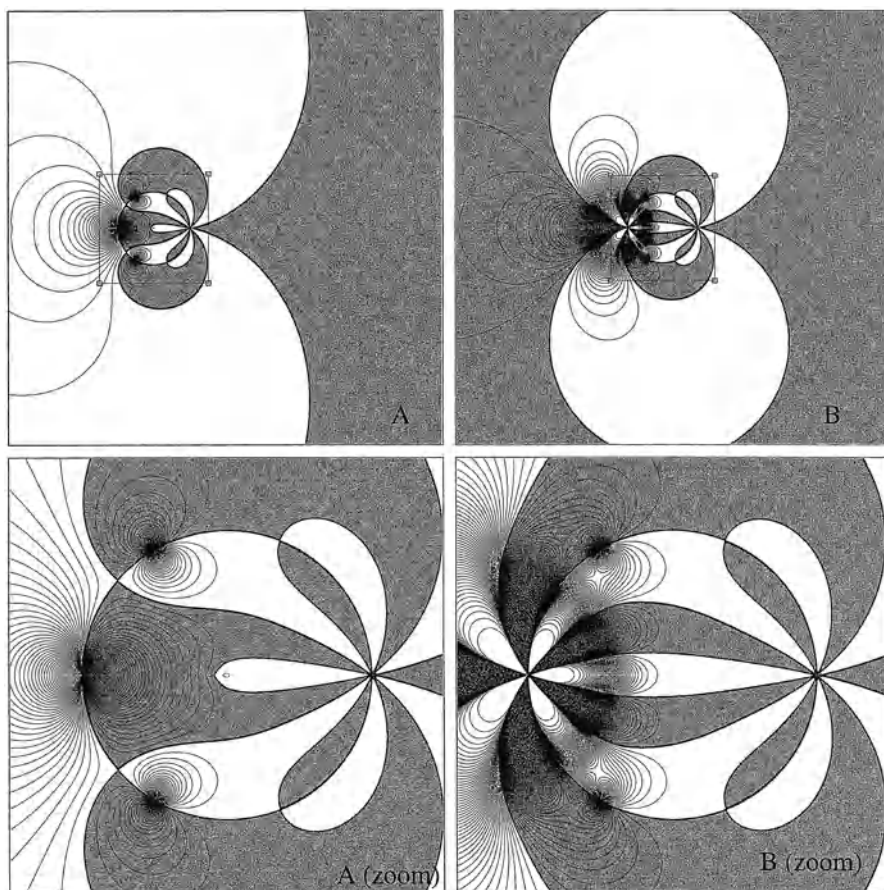


Fig. 4.8. Dual order stars (4.58) and (4.59) for

$$\begin{aligned} \varrho_R(\zeta) &= (\zeta - 1)(\zeta + 1)^5, \quad \varrho(\zeta) = \zeta^6 - 1, \\ \sigma_R(\zeta) &= (251\zeta^6 + 2736\zeta^5 + 6957\zeta^4 + 10352\zeta^3 + 6957\zeta^2 + 2736\zeta + 251)/945 \\ \sigma(\zeta) &= (41\zeta^6 + 216\zeta^5 + 27\zeta^4 + 272\zeta^3 + 27\zeta^2 + 216\zeta + 41)/140 \end{aligned}$$

6. Prove the first Dahlquist barrier by order stars, i.e., prove that stable linear multistep methods satisfy $p \leq k + 2$ (k even) and $p \leq k + 1$ (k odd). Prove also that for methods with optimal order the smallest error constant is assumed by the method with

$$\varrho_R(\zeta) = (\zeta - 1)(\zeta + 1)^{\tilde{k}-1}, \quad (4.60)$$

where $\tilde{k} = k$ (if k is even) and $\tilde{k} = k - 1$ (if k is odd).

Hint. Study the order stars (4.58) (with $\mu = \mu_R$) and (4.59) where $\mu_R = \sigma_R / \varrho_R$ with ϱ_R from (4.60) (see Fig. 4.8 for the case $k = 6$, $p = 8$, $\varrho(\zeta) = \zeta^6 - 1$). You must show that the two order stars in the vicinity of $\zeta = 1$ have the *same* colours. The following observations will help:

- i) The stars in the vicinity of $\zeta = -1$ (produced by the pole $1/(\zeta + 1)^{\tilde{k}-1}$) have *opposite* colours;
- ii) By stability all poles of

$$d_A(\zeta) = \operatorname{Re} \left(\frac{1}{\mu_R(\zeta)} - \frac{1}{\log \zeta} \right), \quad d_B(\zeta) = \operatorname{Re} \left(\frac{1}{\mu_R(\zeta)} - \frac{1}{\mu(\zeta)} \right)$$

lie on or inside the unit circle;

- iii) The boundary curves of A and B cannot cross the unit circle arbitrarily often, since $d_A(e^{i\varphi})$ and $d_B(e^{i\varphi})$ are trigonometric polynomials.

- iv) Study the behaviour of A and B at infinity.

7. Prove the second Dahlquist barrier for linear multistep methods with the help of the order star (4.55).
8. Compute on a computer for an implicit multistep method of order 3 the order star B of (4.18), where $R(\mu)$ is the maximal root of the Milne-Simpson method (1.17). Understand at once the validity of Theorem 4.9.

V.5 Experiments with Multistep Codes

... we know that theory is unable to predict much of what happens in practice at present and software writers need to discover the way ahead by numerical experiment ...

(J.R. Cash, in Aiken 1985)

A comparison of different codes is a notoriously difficult and inexact area ... but there are some clear conclusions that can ...

(J.R. Cash 1983)

This section presents numerical results of multistep codes on precisely the same problems as in Sect. IV.10. These are, in increasing dimension, VDPOL (the van der Pol equation (IV.10.1)), ROBER (the famous Robertson problem (IV.10.2)), OREGO (the Oregonator (IV.10.3)), HIRES (the physiological problem (IV.10.4)), E5 (the badly scaled chemical reaction (IV.10.5)), PLATE ((IV.10.6), a car moving on a plate, the only linear and non autonomous problem), BEAM (the nonlinear elastic beam equation (IV.1.10') with $n = 40$), CUSP (the cusp catastrophe (IV.10.8)), BRUSS (the brusselator (IV.1.6') with one-dimensional diffusion $\alpha = 1/50$ and $n = 500$), and KS (the one-dimensional Kuramoto-Sivashinsky equation (IV.10.11) with $n = 1022$). We have *not* included here the problems BECKDO and BRUSS-2D, since they require a special treatment of the linear algebra routines.

As in Sect. IV.10, the codes have been applied with tolerances

$$Rtol = 10^{-2-m/4} \quad m = 0, 1, 2, \dots$$

and $Atol = Rtol$ (with the exceptions $Atol = 10^{-6} \cdot Rtol$ for OREGO and ROBER, $Atol = 10^{-4} \cdot Rtol$ for HIRES, $Atol = 10^{-3} \cdot Rtol$ for PLATE, and $Atol = 1.7 \cdot 10^{-24}$ for E5). The numerical precisions obtained compared to the CPU times (where all codes are compiled with the same optimization options) are then displayed in Figs. 5.1 and 5.2, again with the symbols representing the required precision $Rtol = 10^{-5}$ displayed in grey tone.

The Codes Used

LSODE — is the “Livermore Solver” of Hindmarsh (1980, 1983). Since we are dealing with stiff equations, we use “stiff” method flags MF=21, 22, 24 or 25, so that the code is based on the Nordsieck representation of the fixed step size BDF methods (see Sections III.6 and III.7). This code emerged from a long development starting with Gear’s DIFSUB in 1971. Its exemplary user interface and ease of application has been a model for much subsequent ODE Software (including ours). Most problems were computed with analytical Jacobian and full linear algebra (MF=21), with the exception of BRUSS and KS (analytical banded Jacobian, MF=24), BEAM (numerical full Jacobian, MF=22), and CUSP (numerical banded Jacobian, MF=25).

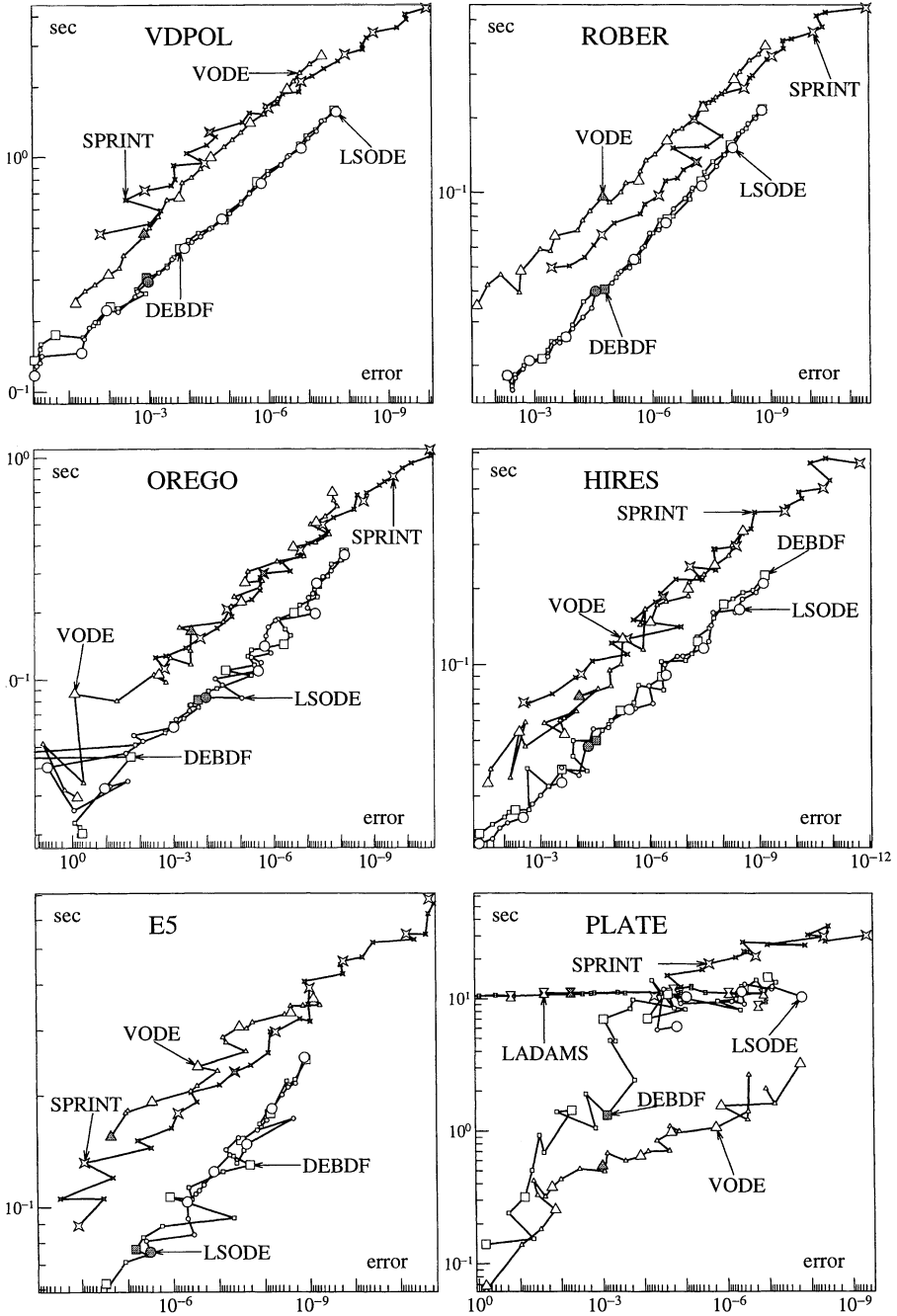


Fig. 5.1. Work-precision diagrams for problems of dimension 2 to 80

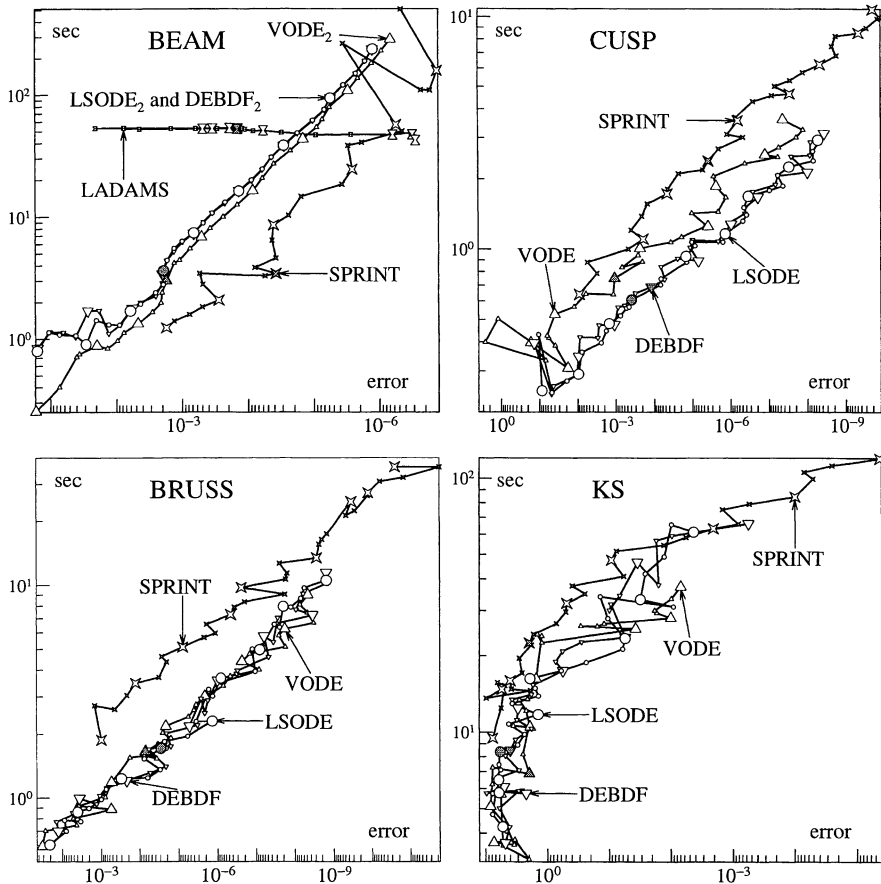


Fig. 5.2. Work-precision diagrams for problems of dimension 80 to 1022

For E5, the code worked correctly only for $Tol \leq 10^{-5}$, for PLATE it was necessary to have $Tol \leq 10^{-7}$. For the BEAM problem, which has eigenvalues on the imaginary axis, it was necessary to restrict the maximal order to 2 because of the lack of A -stability of the higher order BDF methods. The disastrous effect of the allowance of orders higher than 2 can be seen in Fig. 5.3.

DEBDF — this is Shampine & Watts's driver for a modification of the code LSODE and is included in the "DEPAC" family (Shampine & Watts 1979). As is to be expected, it behaves nearly identically to LSODE (see Figs. 5.1 and 5.2). It also requires a restriction of the order for the BEAM problem (see Fig. 5.3).

VODE — is the "Variable-coefficient Ordinary Differential Equation solver" of Brown, Byrne & Hindmarsh (1989). It is based on the EPISODE and EPISODEB packages (see Sect. III.7) which use BDF methods on a non uniform grid (Byrne

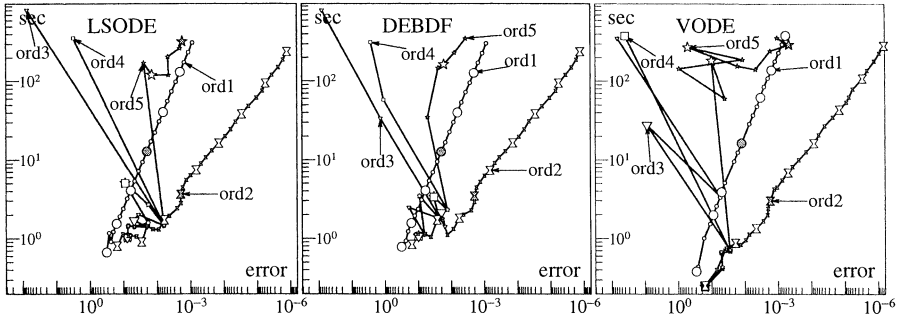


Fig. 5.3. Performance of LSODE, DEBDF and VODE on the BEAM problem with restricted maximal order

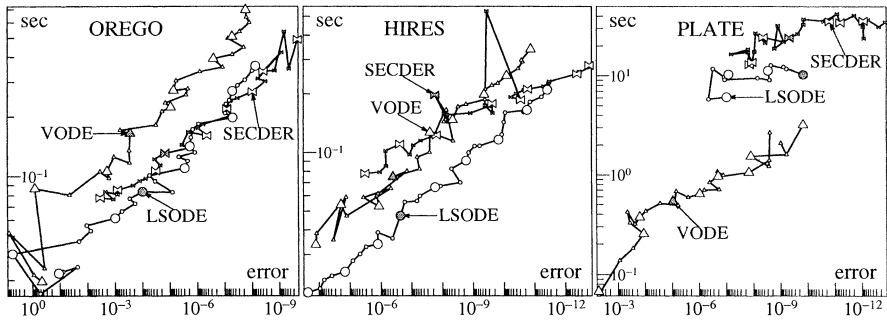


Fig. 5.4. Performance of SECIDER, compared to LSODE and VODE

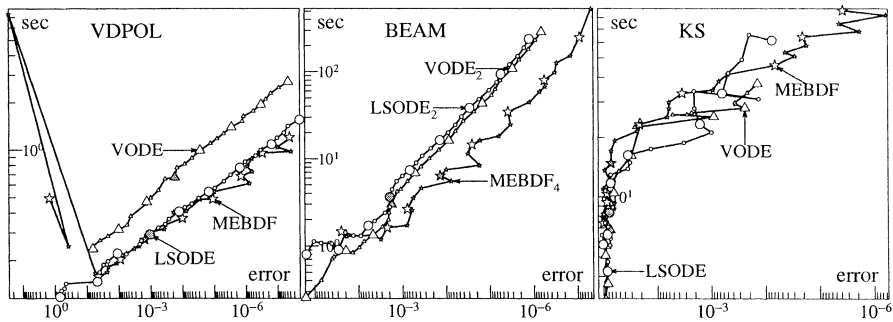


Fig. 5.5. Performance of MEBDF, compared to LSODE and VODE
(for the BEAM problem with restricted maximal order)

& Hindmarsh 1975). The user interface is very similar to that of LSODE; the code again allows selection between full or banded linear algebra and between analytical or numerical Jacobian. The numerical results of VODE (see Figs. 5.1 and 5.2) are very similar for the large problems to those of LSODE and DEBDF, the code is, however, considerably slower on the small problems. For problem E5 this code required a tolerance requirement ($Rtol \leq 10^{-5}$). On the PLATE problem, this code

was by far the best. On the BEAM problem, one has to restrict the maximal order to two (Fig. 5.3).

SPRINT — this package, written by M. Berzins (see Berzins & Furzeland 1985), which has been incorporated into the NAG library (“subchapter D02N”), contains several modules for the step integrator, one of which is SBLEND. This allows us to study the effect of the blended multistep methods (3.15) of Skeel & Kong (1977). It can be seen from Table 3.4 that these methods are A -stable for orders up to 4. We therefore expect them to be much better on the oscillatory BEAM problem. As can be observed in Fig. 5.2 (as well as in Fig. IV.10.8), this code gives excellent results for this problem. An observation of the grey points for $Tol = 10^{-5}$ (Figs. 5.1 and 5.2) shows that the code gives better values than the other multistep codes for a same given tolerance. From time to time, it is fairly slow (e.g., in the PLATE and KS problems).

SECDER — this code, written in 1979 by C.A. Addison (see Addison 1979), implements the SECond DERivative multistep methods (3.7) of Enright. The high order of the methods accompanied with good stability leads us to expect good performance at high tolerances. This has shown to be true (see Fig. 5.4) for OREGO, HIRES and PLATE; however, for the latter it is very slow. We have not used it on the large problems since it has no built-in banded algebra and requires an analytic Jacobian.

MEBDF — this code by Cash & Considine (1992) implements the modified extended BDF methods (see Eq. (3.17.mod) and Table 3.5). Its good performance is shown on selected examples in Fig. 5.5. For the BEAM problem, the code works well if the maximal order is limited to 4.

LADAMS — this is the “Livermore Adams” code, i.e., LSODE with method flag $MF = 10$, included to demonstrate the performance of an *explicit* multistep method on large and/or mildly stiff problems. One can see that it has its chance on several large problems (PLATE, BEAM). It is, when compared to DOPRI5 in Fig. IV.10.8, a good deal slower when f -evaluations are cheap (CUSP), but not on BEAM.

The codes LSODE, DEBDF, VODE and MEBDF can be obtained by sending an electronic mail (e.g., “send lsode.f from odepack”) to “netlib@research.att.com”.

Exercises

1. Do your own experiments and draw your own conclusions for the above problems. The authors will be happy to provide you with drivers and function subroutines.

V.6 One-Leg Methods and G-Stability

... the error analysis is simpler to formulate for one-leg methods than for linear multistep methods. (G. Dahlquist 1975)

The first stability results for *nonlinear* differential equations and multistep methods are fairly old (Liniger 1956, Dahlquist 1963), older than similar studies for Runge-Kutta methods. The great break-through occurred in 1975 (at the Dundee conference) when Dahlquist proposed considering nonlinear problems

$$y' = f(x, y) \quad (6.1)$$

which satisfy a one-sided Lipschitz condition

$$\langle f(x, y) - f(x, z), y - z \rangle \leq \nu \|y - z\|^2 \quad (6.2)$$

or, if the functions are complex-valued,

$$\operatorname{Re} \langle f(x, y) - f(x, z), y - z \rangle \leq \nu \|y - z\|^2 \quad (6.2')$$

(see Sect. IV.12). He also found that the study of nonlinear stability for general multistep methods is simplified, if a related class of methods — the so-called one-leg (multistep) methods — is considered.

One-Leg (Multistep) Methods

... the somewhat crazy name *one-leg methods* ... (G. Dahlquist 1983)

Je ne suis absolument pas capable de traduire “one-leg” en français ... uni-jambiste? (M. Crouzeix, in 1987)

Signor mio, le gru non hanno se non una coscia ed una gamba ... (Boccaccio, Decameròn 1353; quotation suggested by M. Crouzeix)

Suppose that a linear k -step method

$$\sum_{i=0}^k \alpha_i y_{m+i} = h \sum_{i=0}^k \beta_i f(x_{m+i}, y_{m+i}) \quad (6.3)$$

is given, and that the generating polynomials

$$\varrho(\zeta) = \sum_{i=0}^k \alpha_i \zeta^i, \quad \sigma(\zeta) = \sum_{i=0}^k \beta_i \zeta^i \quad (6.4)$$

have real coefficients and no common divisor (see Sect. III.2). We also assume throughout the normalization

$$\sigma(1) = 1. \quad (6.5)$$

Then the associated *one-leg method* is defined by

$$\sum_{i=0}^k \alpha_i y_{m+i} = h f \left(\sum_{i=0}^k \beta_i x_{m+i}, \sum_{i=0}^k \beta_i y_{m+i} \right). \quad (6.6)$$

In this new method, the derivative f is evaluated at one point only, which makes it easier to analyze.

It is, of course, interesting to know how the solutions of the one-leg method (6.6) are related to those of its “multistep twin” (6.3). If the differential equation is linear and autonomous, $y' = Ay$, then both formulas — (6.3) and (6.6) — are identical. For the BDF schemes (1.18) there is in any case only one f -value in the multistep-version, hence the equations (6.3) and (6.6) are the same. For general methods and general nonlinear equations, however, the formulas are *not* identical, but the solutions are related by certain transformations (see Exercise 3). We consider, as an example, the trapezoidal rule, which is a two-leg method,

$$y_{m+1} - y_m = \frac{h}{2} \left(f(x_m, y_m) + f(x_{m+1}, y_{m+1}) \right). \quad (6.7)$$

The corresponding one-leg method is the implicit midpoint rule,

$$y_{m+1} - y_m = h f \left(\frac{x_m + x_{m+1}}{2}, \frac{y_m + y_{m+1}}{2} \right). \quad (6.8)$$

If $\{y_m\}$ is a solution of the one-leg formula (6.8), then

$$\hat{y}_m = \frac{1}{2}(y_m + y_{m+1}), \quad \hat{x}_m = \frac{1}{2}(x_m + x_{m+1})$$

satisfies (6.7). On the other hand, if $\{\hat{y}_m, \hat{x}_m\}$ satisfy (6.7), then

$$y_m = \hat{y}_m - \frac{h}{2} f(\hat{x}_m, \hat{y}_m), \quad x_m = \hat{x}_m - \frac{h}{2}$$

is a solution of (6.8). This relationship has already been extensively exploited in the proof of Theorem IV.15.8.

Existence and Uniqueness

We suppose $\alpha_k \neq 0$ (as always) and $\beta_k \neq 0$ (otherwise the method is explicit). In the case of multistep methods, we write (6.3) in the form

$$y - \eta - h \frac{\beta_k}{\alpha_k} f(x, y) = 0, \quad (6.9)$$

where x is given, η is a vector composed of known quantities, and $y = y_{m+k}$ is the unknown vector. The one-leg Formula (6.6) can also be brought to the form (6.9)

by the transformation $y = \beta_k y_{m+k} + \dots + \beta_0 y_m$, so that all subsequent results on existence and uniqueness will be valid for multistep *and* one-leg methods. To obtain existence results for Eq. (6.9), we replace $h\beta_k/\alpha_k$ by a new “step size” \tilde{h} and obtain nothing else but implicit Euler. All theorems for implicit Runge-Kutta methods (Theorems 14.2, 14.3, and 14.4 of Sect. IV.14) are immediately applicable and give

Theorem 6.1 (Dahlquist 1975). *Let f be continuously differentiable and satisfy (6.2). If*

$$h\nu < \frac{\alpha_k}{\beta_k} \quad (6.10)$$

then the nonlinear equation (6.9) has a unique solution y . \square

Theorem 6.2. *Let y be given by (6.9) and consider a perturbed value \hat{y} satisfying*

$$\hat{y} - \eta - h \frac{\beta_k}{\alpha_k} f(x, \hat{y}) = \delta. \quad (6.11)$$

Under the assumption (6.10) we then have

$$\|\hat{y} - y\| \leq \frac{1}{1 - (\beta_k/\alpha_k)h\nu} \|\delta\|. \quad (6.12)$$

\square

Remark. Theorems IV.14.2, IV.14.3 and IV.14.4 are for much more general methods than just the implicit Euler needed here. The reader who is not interested in the more general case can rewrite the proofs of Sect. IV.14 nearly word for word. Since there is now only one implicit stage, all tensor products disappear and the formulas, but not the ideas of the proof, simplify considerably.

G -Stability

If the differential equation satisfies the one-sided Lipschitz condition (6.2) (or (6.2')) with $\nu = 0$, then the exact solutions are contractive (Lemma IV.12.1). We shall investigate here, which one-leg (multistep) methods then also have contractive solutions. Since the numerical value y_{m+k} depends on all y_{m+k-1}, \dots, y_m , it makes no sense to require $\|y_{m+k} - \hat{y}_{m+k}\| \leq \|y_{m+k-1} - \hat{y}_{m+k-1}\|$ as in the one-step case (Definition IV.12.2). We have to consider the method as a mapping $\mathbb{R}^{n \cdot k} \rightarrow \mathbb{R}^{n \cdot k}$. For this we introduce the notation

$$Y_m = (y_{m+k-1}, \dots, y_m)^T \quad (6.13)$$

and consider inner product norms on $\mathbb{R}^{n \cdot k}$

$$\|Y_m\|_G^2 = \sum_{i=1}^k \sum_{j=1}^k g_{ij} \langle y_{m+i-1}, y_{m+j-1} \rangle, \quad (6.14)$$

where $\langle \cdot, \cdot \rangle$ is the inner product on \mathbb{R}^n used in (6.2) and the k -dimensional matrix

$$G = (g_{ij})_{i,j=1,\dots,k}$$

is assumed to be real, symmetric and positive definite.

Definition 6.3 (Dahlquist 1975). The one-leg method (6.6) is called *G-stable*, if there exists a real, symmetric and positive definite matrix G , such that for two numerical solutions $\{y_m\}$ and $\{\hat{y}_m\}$ we have

$$\|Y_{m+1} - \hat{Y}_{m+1}\|_G \leq \|Y_m - \hat{Y}_m\|_G \quad (6.15)$$

for all step sizes $h > 0$ and for all differential equations satisfying (6.2) or (6.2') with $\nu = 0$.

Since $y' = \lambda y$, $\text{Re } \lambda \leq 0$ satisfies (6.2') with $\nu = 0$, we immediately get

Theorem 6.4. *G-stability implies A-stability.* □

Example 6.5. Consider the 2-step BDF method

$$\frac{3}{2}y_{m+2} - 2y_{m+1} + \frac{1}{2}y_m = hf(x_{m+2}, y_{m+2}). \quad (6.16)$$

We take a second numerical solution $\{\hat{y}_m\}$ and denote its difference to $\{y_m\}$ by $\Delta y_m = y_m - \hat{y}_m$. If we insert (6.16) into our assumption (6.2')

$$\text{Re} \langle f(x_{m+2}, y_{m+2}) - f(x_{m+2}, \hat{y}_{m+2}), y_{m+2} - \hat{y}_{m+2} \rangle \leq 0$$

we obtain

$$E = \text{Re} \left\langle \frac{3}{2}\Delta y_{m+2} - 2\Delta y_{m+1} + \frac{1}{2}\Delta y_m, \Delta y_{m+2} \right\rangle \leq 0. \quad (6.17)$$

The main idea is now to subtract from this inequality a well-chosen quadratic term $\|a_2\Delta y_{m+2} + a_1\Delta y_{m+1} + a_0\Delta y_m\|^2$ in order to bring it to the form required by (6.15). With $\Delta Y_m = (\Delta y_{m+1}, \Delta y_m)^T$ this means that

$$E = \|\Delta Y_{m+1}\|_G^2 - \|\Delta Y_m\|_G^2 + \|a_2\Delta y_{m+2} + a_1\Delta y_{m+1} + a_0\Delta y_m\|^2 \quad (6.18)$$

with a positive definite matrix

$$G = \begin{pmatrix} g_{22} & g_{21} \\ g_{21} & g_{11} \end{pmatrix}.$$

Multiplying out and comparing the coefficients of $\text{Re} \langle \Delta y_i, \Delta y_j \rangle$ in (6.17) and (6.18) gives the six relations

$$\frac{3}{2} = g_{22} + a_2^2, \quad 0 = g_{11} - g_{22} + a_1^2, \quad 0 = -g_{11} + a_0^2, \quad (6.19a)$$

$$-2 = 2g_{21} + 2a_2a_1, \quad \frac{1}{2} = 2a_2a_0, \quad 0 = -2g_{21} + 2a_1a_0. \quad (6.19b)$$

Adding all six equations gives $0 = (a_0 + a_1 + a_2)^2$, so that $a_0 + a_1 + a_2 = 0$. This relation together with (6.19b) determines the a_i as $a_0 = \pm 1/2$, $a_1 = \mp 1$, $a_2 = \pm 1/2$. Inserting this into (6.19) yields the positive definite matrix

$$G = \frac{1}{4} \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix}. \quad (6.20)$$

Since $E \leq 0$ by (6.17), it follows from (6.18) that the 2-step BDF method is G -stable.

An Algebraic Criterion

The algebraic structures of the foregoing computations become much more visible, if we replace formally in (6.17) and (6.18) all

$$\langle \Delta y_{m+i}, \Delta y_{m+j} \rangle \mapsto \zeta^i \omega^j$$

and use

$$2\operatorname{Re} \langle \Delta y_{m+i}, \Delta y_{m+j} \rangle = \langle \Delta y_{m+i}, \Delta y_{m+j} \rangle + \langle \Delta y_{m+j}, \Delta y_{m+i} \rangle.$$

This yields

$$E = \frac{1}{2} (\varrho(\zeta)\sigma(\omega) + \varrho(\omega)\sigma(\zeta)) \quad (6.17')$$

$$E = (\zeta\omega - 1) \sum_{i,j=1}^k g_{ij} \zeta^{i-1} \omega^{j-1} + \left(\sum_{i=0}^k a_i \zeta^i \right) \left(\sum_{j=0}^k a_j \omega^j \right). \quad (6.18')$$

We can now formulate an algebraic criterion which, in a different notation, already appears in Dahlquist (1975).

Theorem 6.6 (Baiocchi & Crouzeix 1989). *Consider a method (ϱ, σ) . If there exists a real, symmetric and positive definite matrix G and real numbers a_0, \dots, a_k , such that*

$$\begin{aligned} & \frac{1}{2} (\varrho(\zeta)\sigma(\omega) + \varrho(\omega)\sigma(\zeta)) \\ &= (\zeta\omega - 1) \sum_{i,j=1}^k g_{ij} \zeta^{i-1} \omega^{j-1} + \left(\sum_{i=0}^k a_i \zeta^i \right) \left(\sum_{j=0}^k a_j \omega^j \right), \end{aligned} \quad (G)$$

then the corresponding one-leg method is G -stable.

Remark. The factor $1/2$ on the left-hand side of (G) is of no significance and can be replaced by any other positive constant, leading to another scaling of the coefficients g_{ij} and a_i .

Proof. We just replace $\zeta^i \omega^j$ by $\langle \Delta y_{m+i}, \Delta y_{m+j} \rangle$ in Eq. (G) and obtain

$$\begin{aligned} \operatorname{Re} \left\langle \sum_{i=0}^k \alpha_i \Delta y_{m+i}, \sum_{j=0}^k \beta_j \Delta y_{m+j} \right\rangle = \\ \|\Delta Y_{m+1}\|_G^2 - \|\Delta Y_m\|_G^2 + \left\| \sum_{i=0}^k a_i \Delta y_{m+i} \right\|^2. \end{aligned} \quad (6.21)$$

We then insert (6.6) and use (6.2') with $\nu = 0$ and obtain the desired estimate $\|\Delta Y_{m+1}\|_G \leq \|\Delta Y_m\|_G$. \square

An interesting question is now, for which methods (ϱ, σ) Condition (6.21) is satisfied. By Theorem 6.4 the method is necessarily A -stable. Is this also sufficient?

The Equivalence of A -Stability and G -Stability

Dahlquist struggled for three years to get the answer, which is

Theorem 6.7 (Dahlquist 1978). *If ϱ and σ have no common divisor, then the method (ϱ, σ) is A -stable if and only if the corresponding one-leg method is G -stable.*

Proof. We follow here the presentation of Baiocchi & Crouzeix (1989). Recall first that A -stability of the method (ϱ, σ) implies

$$\operatorname{Re} \varrho(\zeta) \overline{\sigma(\zeta)} \geq 0 \quad \text{for } |\zeta| \geq 1 \quad (\text{A})$$

(see Sect. V.1). Because of Theorems 6.4 and 6.6 it is sufficient to prove that condition (A) implies the existence of a real, symmetric and positive definite matrix G and real numbers a_0, \dots, a_k such that Property (G) holds. The proof is in three steps:

- a) computation of a_0, \dots, a_k ;
- b) computation of G ;
- c) show that G is positive definite.

a) The term containing the g_{ij} 's in (G) disappears if we put $\omega = 1/\zeta$. We therefore consider the function

$$E(\zeta) = \frac{1}{2} (\varrho(\zeta) \sigma(1/\zeta) + \varrho(1/\zeta) \sigma(\zeta)), \quad (6.22)$$

which is of the form

$$\begin{aligned} E(\zeta) &= c_r \left(\zeta^r + \frac{1}{\zeta^r} \right) + c_{r-1} \left(\zeta^{r-1} + \frac{1}{\zeta^{r-1}} \right) + \dots + c_1 \left(\zeta + \frac{1}{\zeta} \right) + c_0 \\ &= \frac{c_r}{\zeta^r} \prod_{j=1}^{2r} (\zeta - \zeta_j) \end{aligned} \quad (6.23)$$

with some $r \leq k$. Since $E(\zeta) = E(1/\zeta)$, for each root ζ_j of the polynomial $\zeta^r E(\zeta)$ the inverse $1/\zeta_j$ is also a root with the same multiplicity. Therefore there are as many roots *inside* the unit circle as there are *outside*. As to the roots *on* the unit circle, Condition (A) tells us that $E(\zeta) = \operatorname{Re} \varrho(\zeta) \sigma(\bar{\zeta}) \geq 0$ on the unit circle. Therefore, all roots on the unit circle must have *even multiplicity*, half of them we declare “inside” and half of them we declare “outside”. The clever idea is now to collect all roots “outside” the unit circle into a product, so that

$$\begin{aligned} E(\zeta) &= \frac{c_r}{\zeta^r} \prod_{\zeta_j \text{ outside}} (\zeta - \zeta_j) \prod_{\zeta_j \text{ inside}} (\zeta - \zeta_j) \\ &= \frac{c_r}{\zeta^r} \prod_{\zeta_j \text{ outside}} (\zeta - \zeta_j) \prod_{\zeta_j \text{ outside}} \left(\zeta - \frac{1}{\zeta_j} \right) \\ &= K \prod_{\zeta_j \text{ outside}} (\zeta - \zeta_j) \prod_{\zeta_j \text{ outside}} \left(\frac{1}{\zeta} - \zeta_j \right) \end{aligned} \quad (6.24)$$

where K is a constant. But this constant must be non-negative, as can be seen thus: by Condition (A), $E(\zeta)$ is non-negative on the unit circle. The same is true for the function divided by K , since each factor $(e^{i\theta} - \zeta_j)$ from the first product has a complex conjugate brother $(e^{-i\theta} - \bar{\zeta}_j)$ in the second. Therefore $E(\zeta)$ in (6.24) can be factored as

$$E(\zeta) = a(\zeta) \cdot a(1/\zeta) \quad (6.25)$$

where

$$a(\zeta) = \sqrt{K} \prod_{\zeta_j \text{ outside}} (\zeta - \zeta_j) =: \sum_{i=0}^k a_i \zeta^i. \quad (6.26)$$

and step (a) is done.

b) It follows from (6.22) and (6.25) that the polynomial

$$P(\zeta, \omega) = \frac{1}{2} \left(\varrho(\zeta) \sigma(\omega) + \varrho(\omega) \sigma(\zeta) \right) - a(\zeta) a(\omega) \quad (6.27)$$

vanishes when $\zeta\omega - 1 = 0$. It can therefore be written as

$$P(\zeta, \omega) = (\zeta\omega - 1) \sum_{i,j=1}^k g_{ij} \zeta^{i-1} \omega^{j-1}. \quad (6.28)$$

The coefficients g_{ij} are real and satisfy $g_{ij} = g_{ji}$, because $P(\zeta, \omega) = P(\omega, \zeta)$.

c) Looking at (6.28), it appears at first sight a difficult task to prove positive definiteness for the matrix $G = (g_{ij})$ defined there. The crucial idea is the following: choose k (at first arbitrary) complex numbers ζ_1, \dots, ζ_k and replace in (6.28) $\zeta \mapsto \bar{\zeta}_q$, $\omega \mapsto \zeta_r$, which gives together with (6.27)

$$\begin{aligned} b_{qr} &= \sum_{i,j=1}^k \bar{\zeta}_q^{i-1} g_{ij} \zeta_r^{j-1} \\ &= \frac{1}{1 - \bar{\zeta}_q \zeta_r} \left\{ -\frac{1}{2} \left(\varrho(\bar{\zeta}_q) \sigma(\zeta_r) + \varrho(\zeta_r) \sigma(\bar{\zeta}_q) \right) + a(\bar{\zeta}_q) a(\zeta_r) \right\}. \end{aligned} \quad (6.29)$$

Here the b_{qr} are the elements of the matrix

$$B = V^* G V$$

where $V = (\zeta_j^{i-1})$ is a Vandermonde matrix. Thus, we now have to prove that B is positive definite, which appears much easier. First, we develop

$$\frac{1}{1 - \bar{\zeta}_q \zeta_r} = 1 + \bar{\zeta}_q \zeta_r + \bar{\zeta}_q^2 \zeta_r^2 + \bar{\zeta}_q^3 \zeta_r^3 + \dots \quad (6.30a)$$

which converges if

$$|\zeta_q| < 1 \quad q = 1, 2, \dots, k. \quad (6.30b)$$

Next, we require that for all q

$$\varrho(\zeta_q) + \lambda \sigma(\zeta_q) = 0 \quad \text{for some } \lambda > 0. \quad (6.31)$$

With the exception of a finite number of λ 's, the k roots of Eq.(6.31) are all different. A -stability (assumption (A)) implies (6.30b), because $-\lambda$ lies in the interior of the stability domain. Inserting (6.31) and (6.30a) into (6.29) gives, for an arbitrary non-zero vector $\vec{v} = (v_1, \dots, v_k)$,

$$\vec{v}^* B \vec{v} = \sum_{q,r=1}^k \bar{v}_q b_{qr} v_r = \sum_{m=0}^{\infty} \left\{ \left| \sum_{q=1}^k v_q \zeta_q^m a(\zeta_q) \right|^2 + \lambda \left| \sum_{q=1}^k v_q \zeta_q^m \sigma(\zeta_q) \right|^2 \right\},$$

which looks rather positive. This expression cannot be zero for $\vec{v} \neq 0$, because it follows from (6.31) that $\sigma(\zeta_q) \neq 0$ for all q , otherwise ϱ and σ would have a common factor. Therefore $\vec{v}^* B \vec{v} > 0$, thus the matrix B , and consequently the matrix G , is positive definite. \square

It is worth noting that the above proof provides constructive formulas for the matrix G . As an illustration, we again consider the 2-step BDF method (6.16) with generating polynomials

$$\varrho(\zeta) = \frac{3}{2}\zeta^2 - 2\zeta + \frac{1}{2}, \quad \sigma(\zeta) = \zeta^2.$$

The function $E(\zeta)$ (Formula (6.22)) becomes

$$E(\zeta) = \frac{1}{4}\left(\zeta^2 + \frac{1}{\zeta^2}\right) - \left(\zeta + \frac{1}{\zeta}\right) + \frac{3}{2} = \frac{1}{4}(\zeta - 1)^2\left(\frac{1}{\zeta} - 1\right)^2$$

so that $a(\zeta) = \frac{1}{2}(\zeta - 1)^2$. Inserting this into (6.27) gives

$$P(\zeta, \omega) = (\zeta\omega - 1)\left(\frac{5}{4}\zeta\omega - \frac{1}{2}\zeta - \frac{1}{2}\omega + \frac{1}{4}\right),$$

so that $g_{22} = 5/4$, $g_{12} = g_{21} = -1/2$, $g_{11} = 1/4$ is the same as (6.20).

A Criterion for Positive Functions

In the proof of Lemma IV.13.19 we have used the following criterion for positive functions, which is an immediate consequence of the above equivalence result.

Lemma 6.8. *Let $\chi(z) = \alpha(z)/\beta(z)$ be an irreducible rational function with real polynomials $\alpha(z)$ of degree $\leq k-1$ and $\beta(z)$ of degree k . Then $\chi(z)$ is a positive function, i.e.,*

$$\operatorname{Re} \chi(z) > 0 \quad \text{for} \quad \operatorname{Re} z > 0, \quad (6.32)$$

if and only if there exist a real, symmetric and positive definite matrix A and a real, symmetric and non-negative definite matrix B , such that

$$\alpha(z)\beta(w) + \alpha(w)\beta(z) = (z+w) \sum_{i,j=1}^k a_{ij} z^{i-1} w^{j-1} + \sum_{i,j=1}^k b_{ij} z^{i-1} w^{j-1}. \quad (6.33)$$

Proof. The “if”-part follows immediately by putting $w = \bar{z}$ in (6.33). For the “only if”-part we consider the transformations

$$\zeta = \frac{z+1}{z-1}, \quad z = \frac{\zeta+1}{\zeta-1} \quad \text{and} \quad \omega = \frac{w+1}{w-1}, \quad w = \frac{\omega+1}{\omega-1} \quad (6.34)$$

and introduce the polynomials

$$\varrho(\zeta) = \left(\frac{\zeta-1}{2}\right)^k \alpha\left(\frac{\zeta+1}{\zeta-1}\right), \quad \sigma(\zeta) = \left(\frac{\zeta-1}{2}\right)^k \beta\left(\frac{\zeta+1}{\zeta-1}\right).$$

As the transformation (6.34) maps $|\zeta| > 1$ onto the half plane $\operatorname{Re} z > 0$, Condition (6.32) is equivalent to Assumption (A). Therefore, Theorem 6.7 implies the existence of a real, symmetric and positive definite matrix G and of real numbers a_0, \dots, a_k such that

$$\frac{1}{2}(\varrho(\zeta)\sigma(\omega) + \varrho(\omega)\sigma(\zeta)) = (\zeta\omega - 1) \sum_{i,j=1}^k g_{ij} \zeta^{i-1} \omega^{j-1} + \left(\sum_{i=0}^k a_i \zeta^i\right) \left(\sum_{j=0}^k a_j \omega^j\right).$$

Backsubstitution of the old variables yields

$$\begin{aligned}
 & \frac{1}{2}(\alpha(z)\beta(w) + \alpha(w)\beta(z)) \\
 &= 2(z+w) \sum_{i,j=1}^k g_{ij}(z+1)^{i-1}(z-1)^{k-i}(w+1)^{j-1}(w-1)^{k-j} \\
 &+ \left(\sum_{i=0}^k a_i(z+1)^i(z-1)^{k-i} \right) \left(\sum_{j=0}^k a_j(w+1)^j(w-1)^{k-j} \right).
 \end{aligned} \tag{6.35}$$

Rearranging into powers of z and w gives Eq. (6.33). Since the polynomials $(z+1)^{i-1}(z-1)^{k-i}$ for $i = 1, \dots, k$ are linearly independent, the resulting matrix A is positive definite. The coefficient of $z^k w^k$ in the second term of the right-hand side of (6.35) must vanish, because the degree of $\alpha(z)$ is at most $k-1$. We remark that the matrix B of this construction is only of rank 1. \square

Error Bounds for One-Leg Methods

We shall apply the stability results of this section to derive bounds for the global error of one-leg methods. For a differential equation (6.1) with exact (smooth) solution $y(x)$ it is natural to define the discretization error of (6.6) as

$$\delta_{OL}(x) = \sum_{i=0}^k \alpha_i y(x+ih) - hf\left(x+\beta h, \sum_{i=0}^k \beta_i y(x+ih)\right) \tag{6.36}$$

with $\beta = \sigma'(1) = \sum i\beta_i$. For the BDF methods we have $\sum_i \beta_i y(x+ih) = y(x+\beta h)$, so that (6.36) equals

$$\delta_D(x) = \sum_{i=0}^k \alpha_i y(x+ih) - hy'(x+\beta h), \tag{6.37}$$

the so-called *differentiation error* of the method. For methods which do not satisfy $\sum_i \beta_i y(x+ih) = y(x+\beta h)$, the right hand side of (6.36) may become very large for stiff problems, even if the derivatives of the solution are bounded by a constant of moderate size. In this case, the expression (6.36) is not a suitable quantity for error estimates. Dahlquist (1983) proposed considering in addition to $\delta_D(x)$ also the *interpolation error*

$$\delta_I(x) = \sum_{i=0}^k \beta_i y(x+ih) - y(x+\beta h). \tag{6.38}$$

For nonstiff problems (with bounded derivatives of f) these two error expressions are related to $\delta_{OL}(x)$ by

$$\delta_{OL}(x) = \delta_D(x) - h \frac{\partial f}{\partial y}(x, y(x)) \delta_I(x) + \mathcal{O}(h \|\delta_I(x)\|^2).$$

Taylor expansion of (6.37) and (6.38) shows that

$$\delta_D(x) = \mathcal{O}(h^{p_D+1}), \quad \delta_I(x) = \mathcal{O}(h^{p_I+1}), \quad (6.39)$$

where the optimal orders p_D and p_I are determined by certain algebraic conditions (see Exercise 1a). From $\beta = \sigma'(1)$ we always have $p_I \geq 1$ and from the consistency conditions it follows that $p_D \geq 1$. However, the orders p_D and p_I may be significantly smaller than the order of the corresponding multistep method (Exercise 1). The constants in the $\mathcal{O}(\dots)$ -terms of (6.39) depend only on bounds for a certain derivative of the solution, but not on the stiffness of the problem.

Using $\delta_D(x)$ and $\delta_I(x)$ it is possible to interpret the exact solution of (6.1) as the solution of the following perturbed one-leg formula

$$\sum_{i=0}^k \alpha_i y(x+ih) - \delta_D(x) = hf\left(x + \beta h, \sum_{i=0}^k \beta_i y(x+ih) - \delta_I(x)\right). \quad (6.40)$$

The next lemma, which extends results of Dahlquist (1975) and of Nevanlinna (1976), investigates the influence of perturbations to the solution of a one-leg method.

Lemma 6.9. *Consider, in addition to the one-leg method (6.6), the perturbed formula*

$$\sum_{i=0}^k \alpha_i \widehat{y}_{m+i} - \delta_m = hf\left(x_m + \beta h, \sum_{i=0}^k \beta_i \widehat{y}_{m+i} - \varepsilon_m\right). \quad (6.41)$$

Suppose that the condition (6.2') holds for the differential equation (6.1) and that the method is G -stable. Then the differences

$$\Delta y_j = \widehat{y}_j - y_j, \quad \Delta Y_m = (\Delta y_{m+k-1}, \dots, \Delta y_m)^T$$

satisfy in the norm (6.14)

$$\|\Delta Y_{m+1}\|_G \leq (1 + ch\nu) \|\Delta Y_m\|_G + C(\|\delta_m\| + \|\varepsilon_m\|) \quad \text{for } 0 < h\nu \leq \text{Const.}$$

The constants c , C , and Const depend only on the method, not on the differential equation. If $\nu \leq 0$ we have

$$\|\Delta Y_{m+1}\|_G \leq \|\Delta Y_m\|_G + C(\|\delta_m\| + \|\varepsilon_m\|) \quad \text{for all } h > 0.$$

Proof. We shall make the additional assumption that f is continuously differentiable. A direct proof without this assumption is possible, but leads to a quadratic inequality for $\|\Delta Y_{m+1}\|_G$.

The idea is to subtract (6.6) from (6.41) and to use

$$\begin{aligned} f(x_m + \beta h, \sum \beta_i \widehat{y}_{m+i} - \varepsilon_m) - f(x_m + \beta h, \sum \beta_i y_{m+i}) \\ = J_m(\sum \beta_i \Delta y_{m+i} - \varepsilon_m) \end{aligned}$$

where

$$J_m = \int_0^1 \frac{\partial f}{\partial y} \left(x_m + \beta h, t \sum \beta_i y_{m+i} + (1-t) \left(\sum \beta_i \hat{y}_{m+i} - \varepsilon_m \right) \right) dt.$$

This yields

$$\sum_{i=0}^k \alpha_i \Delta y_{m+i} = h J_m \sum_{i=0}^k \beta_i \Delta y_{m+i} + \delta_m - h J_m \varepsilon_m.$$

Computing Δy_{m+k} from this relation gives

$$\Delta y_{m+k} = \Delta z_{m+k} + (\alpha_k - \beta_k h J_m)^{-1} (\delta_m - h J_m \varepsilon_m) \quad (6.42)$$

where Δz_{m+k} is defined by

$$\sum_{i=0}^k \alpha_i \Delta z_{m+i} = h J_m \sum_{i=0}^k \beta_i \Delta z_{m+i} \quad (6.43)$$

and $\Delta z_j = \Delta y_j$ for $j < m+k$. By our assumption (6.2') the matrix J_m satisfies the one-sided Lipschitz condition $\operatorname{Re} \langle J_m u, u \rangle \leq \nu \|u\|^2$ (see Exercise 6 of Sect. I.10). Taking the scalar product of (6.43) with $\sum \beta_i \Delta z_{m+i}$ and using (6.21) we thus obtain in the notation of (6.13)

$$\begin{aligned} \|\Delta Z_{m+1}\|_G^2 - \|\Delta Z_m\|_G^2 &\leq c_0 h \nu \left\| \sum \beta_i \Delta z_{m+i} \right\|^2 \\ &\leq c_1 h \nu (\|\Delta Z_{m+1}\|_G + \|\Delta Z_m\|_G)^2 \end{aligned}$$

(the second inequality is only valid for $\nu \geq 0$; for negative values of ν we replace ν by 0 in (6.2')). A division by $\|\Delta Z_{m+1}\|_G + \|\Delta Z_m\|_G$ then leads to the estimate

$$\|\Delta Z_{m+1}\|_G \leq (1 + ch\nu) \|\Delta Z_m\|_G. \quad (6.44)$$

With the help of von Neumann's theorem (Sect. IV.11) the second term of (6.42) can be bounded by $\operatorname{Const} (\|\delta_m\| + \|\varepsilon_m\|)$. Inserting this and (6.44) into (6.42) yields the desired estimate. \square

The above lemma allows us to derive a convergence result for one-leg methods, which is related to B -convergence for Runge-Kutta methods.

Theorem 6.10. *Consider a G -stable one-leg method with differentiation order $p_D \geq p$ and interpolation order $p_I \geq p-1$. Suppose that the differential equation satisfies the one-sided Lipschitz condition (6.2'). Then there exists $C_0 > 0$ such that for $h\nu \leq C_0$*

$$\|y_m - y(x_m)\| \leq C \max_{0 \leq j < k} \|y_j - y(x_j)\| + M h^p. \quad (6.45)$$

The constant C depends on the method and, for $\nu > 0$, on the length $x_m - x_0$ of the integration interval; the constant M depends in addition on bounds for the p -th and $(p+1)$ -th derivative of the exact solution.

Proof. A direct application of Lemma 6.9 to Eq. (6.40) yields the desired error bounds only for $p_I \geq p$. Following Hundsdorfer & Steiner (1991) we therefore introduce $\widehat{y}(x) = y(x) - \delta_I(x)$, so that (6.40) becomes

$$\sum_{i=0}^k \alpha_i \widehat{y}(x + ih) - \widehat{\delta}(x) = hf(x + \beta h, \sum_{i=0}^k \beta_i \widehat{y}(x + ih) - \widehat{\varepsilon}(x)), \quad (6.46)$$

where

$$\widehat{\delta}(x) = \delta_D(x) - \sum_{i=0}^k \alpha_i \delta_I(x + ih), \quad \widehat{\varepsilon}(x) = \delta_I(x) - \sum_{i=0}^k \beta_i \delta_I(x + ih). \quad (6.47)$$

Using $\varrho(1) = 0$ and $\sigma(1) = 1$, Taylor expansion of these functions yields

$$\|\widehat{\delta}(x)\| + \|\widehat{\varepsilon}(x)\| \leq C_1 h^p \int_x^{x+kh} \|y^{(p+1)}(\zeta)\| d\zeta.$$

We thus can apply Lemma 6.9 to (6.46) and obtain

$$\|\Delta Y_{m+1}\|_G \leq (1 + ch\nu) \|\Delta Y_m\|_G + M_1 h^{p+1}$$

where $\Delta y_j = \widehat{y}(x_j) - y_j$. Using $(1 + ch\nu)^j \leq \exp(c\nu(x_j - x_0))$, a simple induction argument gives

$$\|\Delta Y_{m+1}\|_G \leq C \|\Delta Y_0\|_G + M h^p.$$

The statement now follows from the equivalence of norms

$$d_0 \|\Delta Y_0\|_G \leq \max_{0 \leq j < k} \|\Delta y_j\| \leq d_1 \|\Delta Y_0\|_G,$$

from the estimate $\|y_m - y(x_m)\| \leq \|y_m - \widehat{y}(x_m)\| + \|\delta_I(x_m)\|$, and from the fact that $\|\delta_I(x_m)\| = \mathcal{O}(h^p)$. \square

Convergence of A -Stable Multistep Methods

An interesting equivalence relation between one-leg and linear multistep methods is presented in Dahlquist (1975) (see Exercise 3). This allows us to translate the above convergence result into a corresponding one for multistep methods (Hundsdorfer & Steiner 1991). A different and more direct approach will be presented in Sect. V.8 (Theorem 8.2).

We consider the linear multistep method

$$\sum_{i=0}^k \alpha_i \widehat{y}_{m+i} = h \sum_{i=0}^k \beta_i f(\widehat{x}_{m+i}, \widehat{y}_{m+i}). \quad (6.48)$$

We let $x_m = \hat{x}_m - \beta h$, so that $\sum_{i=0}^k \beta_i x_{m+i} = \hat{x}_m$, and, in view of Eq. (6.54), we define $\{y_0, y_1, \dots, y_{2k-1}\}$ as the solution of the linear system

$$\sum_{i=0}^k \beta_i y_{j+i} = \hat{y}_j, \quad \sum_{i=0}^k \alpha_i y_{j+i} = h f(\hat{x}_j, \hat{y}_j), \quad j = 0, \dots, k-1. \quad (6.49)$$

This system is uniquely solvable, because the polynomials $\varrho(\zeta)$ and $\sigma(\zeta)$ are relatively prime. With these starting values we define $\{y_m\}$ as solution of the one-leg relation (for $m \geq k$)

$$\sum_{i=0}^k \alpha_i y_{m+i} = h f\left(\sum_{i=0}^k \beta_i x_{m+i}, \sum_{i=0}^k \beta_i y_{m+i}\right). \quad (6.50)$$

By the second relation of (6.49), Eq. (6.50) holds for all $m \geq 0$. Consequently (Exercise 3a) the expression $\sum_{i=0}^k \beta_i y_{m+i}$ is a solution of the multistep method (6.48). Because of (6.49) and the uniqueness of the numerical solution this gives

$$\sum_{i=0}^k \beta_i y_{m+i} = \hat{y}_m \quad \text{for all } m \geq 0. \quad (6.51)$$

This relation leads to a proof of the following result.

Theorem 6.11. *Consider an A-stable linear multistep method of order p . Suppose the differential equation satisfies (6.2'). Then there exists $C_0 > 0$ such that for $h\nu \leq C_0$,*

$$\|\hat{y}_m - y(\hat{x}_m)\| \leq C \left(\max_{0 \leq j < k} \|\hat{y}_j - y(\hat{x}_j)\| + h \max_{0 \leq j < k} \|f(\hat{x}_j, \hat{y}_j) - y'(\hat{x}_j)\| \right) + M h^p.$$

The constants C and M are as in Theorem 6.10.

Proof. By Theorem 6.7, A-stability implies G -stability of the corresponding one-leg method. Further, Taylor expansion of (6.37) and (6.38) shows that $p_D \geq \min(p, 2)$ and $p_I \geq 1$. Since $p \leq 2$ by Dahlquist's second barrier, all assumptions of Theorem 6.10 are verified. The one-leg solution $\{y_m\}$ thus satisfies (6.45). In order to estimate $\|y_j - y(x_j)\|$ for $j < k$ we subtract the definitions of $\delta_D(x)$ and $\delta_I(x)$ from (6.48) and obtain

$$\begin{aligned} \sum_{i=0}^k \beta_i (y_{j+i} - y(x_{j+i})) &= \hat{y}_j - y(\hat{x}_j) - \delta_I(x_j) \\ \sum_{i=0}^k \alpha_i (y_{j+i} - y(x_{j+i})) &= h f(\hat{x}_j, \hat{y}_j) - h y'(\hat{x}_j) - \delta_D(x_j). \end{aligned}$$

Solving these relations for $y_j - y(x_j)$ yields

$$\begin{aligned} \max_{0 \leq j < k} \|y_j - y(x_j)\| \\ \leq C_0 \left(\max_{0 \leq j < k} \|\hat{y}_j - y(\hat{x}_j)\| + h \max_{0 \leq j < k} \|f(\hat{x}_j, \hat{y}_j) - y'(\hat{x}_j)\| \right) + M_0 h^p. \end{aligned}$$

This proves the statement, because by (6.51)

$$\|\hat{y}_m - y(\hat{x}_m)\| \leq \sum_{i=0}^k |\beta_i| \|y_{m+i} - y(x_{m+i})\| + \|\delta_I(x_m)\|. \quad \square$$

Exercises

1. a) Prove that the one-leg method (6.6) satisfies (6.39) iff

$$\sum_{i=0}^k \alpha_i i^q = q\beta^{q-1} \quad \text{for } q = 0, 1, \dots, p_D \quad (6.52)$$

$$\sum_{i=0}^k \beta_i i^q = \beta^q \quad \text{for } q = 0, \dots, p_I. \quad (6.53)$$

Compare this result with Theorem III.2.4.

- b) Compute the orders p_D and p_I for the Adams methods.
2. a) Show that the one-leg method (6.6) can be written in the form of a general linear method (Sect. III.8).
- b) Prove that the order of convergence p of this method is given by

$$p = \min(p_D, p_I + 1)$$

with p_D, p_I defined in (6.39).

- c) The order of a one-leg method is never larger than the order of the corresponding multistep method.
3. (Dahlquist 1975).
- a) Let $\{y_m\}$ and $\{x_m = x_0 + mh\}$ satisfy the (one-leg) difference relation (6.6); then

$$\hat{y}_m = \sum_{j=0}^k \beta_j y_{m+j}, \quad \hat{x}_m = \sum_{j=0}^k \beta_j x_{m+j} \quad (6.54)$$

satisfy the (linear multistep) difference relation (6.3).

- b) Conversely, let

$$P(\zeta) = \sum_{j=0}^{k-1} a_j \zeta^j, \quad Q(\zeta) = \sum_{j=0}^{k-1} b_j \zeta^j$$

be such that $P(\zeta)\sigma(\zeta) - Q(\zeta)\varrho(\zeta) = \zeta^l$ for some integer l ($0 \leq l \leq k$), then

$$y_{m+l} = \sum_{j=0}^{k-1} a_j \hat{y}_{m+j} - h \sum_{j=0}^{k-1} b_j f(\hat{x}_{m+j}, \hat{y}_{m+j})$$

$$x_{m+l} = \sum_{j=0}^{k-1} a_j \hat{x}_{m+j} - h \sum_{j=0}^{k-1} b_j$$

satisfy (6.6), whenever $\{\hat{y}_m\}$ and $\{\hat{x}_m\}$ are a solution of (6.3).

Hint for a). Multiply (6.6) by β_j , replace m by $m+j$, sum from $j=0$ to $j=k$, and interchange the summations.

4. *One-leg collocation methods* (Dahlquist 1983).

a) For a given β there exists a unique k -step one-leg method with $p_D = k$ and $p_I = k$.

b) This one-leg method is of order $p = k+1$ iff

$$\sum_{i=0}^k \frac{1}{(\beta - i)} = 0.$$

c) Discuss numerically the zero-stability of these methods.

5. (proposed by M. Crouzeix). a) Let $R(z) = P(z)/Q(z)$ be an irreducible rational function where $\deg P \leq k$, $\deg Q \leq k$. Show that $R(z)$ is A -stable, if and only if there exist polynomials $\alpha_i(z)$, $\beta(z)$ with real coefficients and with $\deg \alpha_i \leq k-1$, $\deg \beta \leq k$, such that

$$Q(z)Q(w) - P(z)P(w) = -(z+w) \sum_{i=1}^k \alpha_i(z)\alpha_i(w) + \beta(z)\beta(w). \quad (6.55)$$

b) Use this characterization to give a new proof of von Neumann's theorem (Corollary IV.11.3).

Hint. Part (a) can be proved along the lines of the proofs of Theorem 6.7 and Lemma 6.8. Remark that (6.55) reduces to the E -polynomial (IV.3.8) if $z = iy$ and $w = -iy$. For the proof of (b), deduce from (6.55) the identity

$$\|Q(A)u\|^2 - \|P(A)u\|^2 = - \sum_{i=1}^k \operatorname{Re} \langle \alpha_i(A)u, A\alpha_i(A)u \rangle + \|\beta(A)u\|^2.$$

V.7 Convergence for Linear Problems

Theorems 6.10 and 6.11 give satisfactory convergence results for G -stable one-leg methods and A -stable multistep methods. But there are only few such methods and their highest order is two (Theorem 1.4). It is therefore interesting to relax the requirement of A -stability and to investigate higher-order multistep and one-leg methods. This section is devoted to linear stiff problems, while Sect. V.8 will treat non-linear problems.

We shall describe two different approaches for convergence results. One is with the help of the discrete variation of constants formula and shall be given at the end of this section (see Lemma 7.9 and Theorem 7.10 below). The other possibility is based on a formulation as a one-step method and on the use of the Kreiss matrix theorem.

Difference Equations for the Global Error

Most of the difficulties can already be seen by studying the one-dimensional problem of Prothero and Robinson

$$y' = \lambda y + g(x), \quad y(x_0) = y_0. \quad (7.1)$$

We assume $\operatorname{Re} \lambda \leq 0$ and the solution $y(x)$ to be smooth in the sense that sufficiently many derivatives are bounded independently of the stiffness parameter λ .

Applying a *linear multistep method* to (7.1) yields

$$\sum_{i=0}^k \alpha_i y_{m+i} = h\lambda \sum_{i=0}^k \beta_i y_{m+i} + h \sum_{i=0}^k \beta_i g(x_{m+i}). \quad (7.2)$$

The global error

$$e_m = y_m - y(x_m) \quad (7.3)$$

is seen to satisfy the difference relation

$$\sum_{i=0}^k (\alpha_i - h\lambda\beta_i) e_{m+i} = -\delta_{LM}(x_m) \quad (7.4)$$

with

$$\delta_{LM}(x) = \sum_{i=0}^k \alpha_i y(x+ih) - h \sum_{i=0}^k \beta_i y'(x+ih) \quad (7.5)$$

(to be compared with Formula III.2.3). We observe that the right-hand side of (7.4) is independent of the stiffness (i.e., of λ). Further, if the classical order of the method is p , then $\delta_{LM}(x) = \mathcal{O}(h^{p+1})$.

If we apply the method in its *one-leg* version, we obtain

$$\sum_{i=0}^k \alpha_i y_{m+i} = h\lambda \sum_{i=0}^k \beta_i y_{m+i} + hg(x_m + \beta h), \quad (7.6)$$

where $\sum \beta_i = 1$ and $\sum \beta_i i = \beta$. In this case the global error $e_m = y_m - y(x_m)$ satisfies

$$\sum_{i=0}^k (\alpha_i - h\lambda\beta_i) e_{m+i} = h\lambda\delta_I(x_m) - \delta_D(x_m) \quad (7.7)$$

with $\delta_D(x)$ and $\delta_I(x)$ defined in (6.37) and (6.38), respectively. Unless $\delta_I(x) = 0$ (which is the case for the BDF methods), Eq. (7.7) is disappointing, because its right-hand side becomes large in the stiff case ($h\lambda \rightarrow \infty$).

In order to overcome this difficulty, Dahlquist (1983) proposes that one consider instead of $e_m = y_m - y(x_m)$ the quantities

$$e_m^* = \sum_{i=0}^k \beta_i y_{m+i} - y(x_m + \beta h) \quad (7.8)$$

(“... a more adequate measure of the global error than the customary one ...”, Dahlquist 1983). Replacing m by $m+j$ in (7.6), multiplying by β_j and summing up gives the error formula

$$\sum_{i=0}^k (\alpha_i - h\lambda\beta_i) e_{m+i}^* = -\delta_{LM}(x_m + \beta h) \quad (7.9)$$

with $\delta_{LM}(x)$ of (7.5). This difference relation now has the same strength as (7.4).

It has been pointed out by Hundsdorfer & Steiner (1991) that we usually get better error estimates for one-leg methods by considering $\widehat{e}_m = e_m + \delta_I(x_m)$. We then have

$$\sum_{i=0}^k (\alpha_i - h\lambda\beta_i) \widehat{e}_{m+i} = h\lambda\widehat{\varepsilon}(x_m) - \widehat{\delta}(x_m) \quad (7.10)$$

with $\widehat{\varepsilon}(x)$ and $\widehat{\delta}(x)$ given by (6.47). Observe that $\widehat{\varepsilon}(x) = \mathcal{O}(h^{p_I+2})$ and $\widehat{\delta}(x) = \mathcal{O}(h^{\min(p_D+1, p_I+2)})$.

Formulation as a One-Step Method. The error relations (7.4), (7.7), (7.9), and (7.10) are all of the form

$$\sum_{i=0}^k (\alpha_i - h\lambda\beta_i) e_{m+i} = \delta_h(x_m). \quad (7.11)$$

In order to estimate e_m it is convenient to introduce, as in Sect. III.4, the vector

$$E_m = (e_{m+k-1}, \dots, e_{m+1}, e_m)^T, \quad (7.12)$$

the companion matrix

$$C(\mu) = \begin{pmatrix} c_{k-1}(\mu) & \dots & c_1(\mu) & c_0(\mu) \\ 1 & & & \\ & \ddots & & \\ & & 1 & 0 \end{pmatrix}, \quad c_j(\mu) = -\frac{\alpha_j - \mu\beta_j}{\alpha_k - \mu\beta_k} \quad (7.13)$$

and

$$\Delta_m = (\delta_h(x_m)/(\alpha_k - \mu\beta_k), 0, \dots, 0)^T, \quad \mu = h\lambda. \quad (7.14)$$

Then, Eq. (7.11) becomes

$$E_{m+1} = C(h\lambda)E_m + \Delta_m, \quad (7.15)$$

which leads to

$$E_{m+1} = C(h\lambda)^{m+1}E_0 + \sum_{j=0}^m C(h\lambda)^{m-j}\Delta_j. \quad (7.16)$$

To estimate E_{m+1} we have to bound the powers of $C(h\lambda)$ uniformly in $h\lambda$. This is the subject of the next subsection.

The Kreiss Matrix Theorem

Als Fakultätsopponent für meine Stockholmer Dissertation brachte Dr. G. Dahlquist die Frage der Stabilitätsdefinition zur Sprache.

(H.-O. Kreiss 1962)

The following Theorem of Kreiss (1962) is a powerful tool for proving uniform power boundedness of an arbitrary family of matrices.

Theorem 7.1 (Kreiss 1962). *Let \mathcal{F} be a family of $k \times k$ matrices A . Then the “power condition”*

$$\|A^n\| \leq M \quad \text{for } n = 0, 1, 2, \dots \quad \text{and} \quad A \in \mathcal{F} \quad (P)$$

is equivalent to the “resolvent condition”

$$\|(A - zI)^{-1}\| \leq \frac{C}{|z| - 1} \quad \text{for } |z| > 1 \quad \text{and} \quad A \in \mathcal{F}. \quad (R)$$

Remark. The difficult step is to prove that (R) implies (P) . Several mathematicians contributed to a better understanding of this result (Richtmyer & Morton 1967, Tadmor 1981). LeVeque & Trefethen (1984) have given a marvellous version of the proof; the best we can do is to copy it nearly word for word:

Proof. Necessity. If (P) is true, the eigenvalues of A lie within the closed unit disk and therefore $(A - zI)^{-1}$ exists for $|z| > 1$. Moreover,

$$\|(A - zI)^{-1}\| = \left\| \sum_{n=0}^{\infty} A^n z^{-n-1} \right\| \leq M \sum_{n=0}^{\infty} |z|^{-n-1} = \frac{M}{|z| - 1}, \quad (7.17)$$

so that (R) holds with $C = M$.

Sufficiency. Assume condition (R) , so that all eigenvalues of A lie inside the closed unit disk. The matrix A^n can then be written in terms of the resolvent by means of a Cauchy integral (see Exercise 1)

$$A^n = \frac{1}{2\pi i} \int_{\Gamma} z^n (zI - A)^{-1} dz, \quad (7.18)$$

where the contour of integration is, for example, a circle of radius $\varrho > 1$ centred at the origin. Let u and v be arbitrary unit vectors, i.e., $\|u\| = \|v\| = 1$. Then,

$$v^* A^n u = \frac{1}{2\pi i} \int_{\Gamma} z^n q(z) dz \quad \text{with} \quad q(z) = v^* (zI - A)^{-1} u.$$

Integration by parts gives

$$v^* A^n u = \frac{-1}{2\pi i(n+1)} \int_{\Gamma} z^{n+1} q'(z) dz.$$

Now fix as contour of integration the circle of radius $\varrho = 1 + 1/(n+1)$. On this path one has $|z^{n+1}| \leq e$, and therefore

$$|v^* A^n u| \leq \frac{e}{2\pi(n+1)} \int_{\Gamma} |q'(z)| |dz|. \quad (7.19)$$

By Cramer's rule, $q(z)$ is a rational function of degree k . Applying Lemma 7.2 below, the integral in (7.19) is bounded by $4\pi k$ times the supremum of $|q'(z)|$ on Γ , and by (R) this supremum is at most $(n+1)C$. Hence

$$|v^* A^n u| \leq 2ekC.$$

Since $\|A^n\|$ is the supremum of $|v^* A^n u|$ over all unit vectors u and v , this proves the estimate (P) with $M = 2ekC$. \square

The above proof used the following lemma, which relates the arc length of a rational function on a circle to its maximum value. For the case of a polynomial of degree k the result is a corollary of Bernstein's inequality $\sup_{|z|=1} |q'(z)| \leq k \sup_{|z|=1} |q(z)|$ (see e.g., Marden 1966).

Lemma 7.2. Let $q(z) = p(z)/r(z)$ be a rational function with $\deg p \leq k$, $\deg r \leq k$ and suppose that no poles lie on the circle $\Gamma : |z| = \varrho$. Then

$$\int_{\Gamma} |q'(z)| |dz| \leq 4\pi k \sup_{|z|=\varrho} |q(z)|. \quad (7.20)$$

Proof. Replacing $q(z)$ by $q(\varrho z)$ we may assume without loss of generality that $\varrho = 1$. With the parametrization e^{it} of Γ we introduce

$$\gamma(t) = q(e^{it}), \quad \gamma'(t) = ie^{it} q'(e^{it})$$

so that

$$\gamma'(t) = |q'(e^{it})| \cdot e^{ig(t)} \quad \text{with} \quad g(t) = \arg(\gamma'(t)).$$

Integration by parts now yields

$$\begin{aligned} \int_{\Gamma} |q'(z)| |dz| &= \int_0^{2\pi} |q'(e^{it})| dt = \int_0^{2\pi} \gamma'(t) e^{-ig(t)} dt \\ &= i \int_0^{2\pi} \gamma(t) g'(t) e^{-ig(t)} dt \leq \sup |\gamma(t)| \cdot \int_0^{2\pi} |g'(t)| dt. \end{aligned}$$

It remains to prove that the total variation of g , i.e., $\text{TV}[g] = \int_0^{2\pi} |g'(t)| dt$, can be bounded by $4\pi k$. To prove this, note that $zq'(z)$ is a rational function of degree $(2k, 2k)$ and can be written as a product

$$zq'(z) = \prod_{j=1}^{2k} \frac{a_j z + b_j}{c_j z + d_j}.$$

This implies for $z = e^{it}$

$$g(t) = \arg(izq'(z)) = \frac{\pi}{2} + \sum_{j=1}^{2k} \arg\left(\frac{a_j z + b_j}{c_j z + d_j}\right).$$

Since the Möbius transformation $(az + b)/(cz + d)$ maps the unit circle to some other circle, the total variation of $\arg((az + b)/(cz + d))$ is at most 2π . Consequently,

$$\text{TV}[g] \leq \sum_{j=1}^{2k} \text{TV}\left[\arg\left(\frac{a_j z + b_j}{c_j z + d_j}\right)\right] \leq 4\pi k. \quad \square$$

Remark. It has been conjectured by LeVeque & Trefethen (1984) that the bound (7.20) is valid with a factor 2π instead of 4π . This conjecture has been proved to be true by Spijker (1991).

Some Applications of the Kreiss Matrix Theorem

Following Dahlquist, Mingyou & LeVeque (1983) we now obtain some results on the uniform power boundedness of the matrix $C(\mu)$, defined in (7.13), with the help of the Kreiss matrix theorem. Similar results were found independently by Crouzeix & Raviart (1980) and Gekeler (1979, 1984).

Lemma 7.3. *Let $S \subset \overline{\mathbb{C}}$ denote the stability region of a method (ϱ, σ) . If S is closed in $\overline{\mathbb{C}}$, then there exists a constant M such that*

$$\|C(\mu)^n\| \leq M \quad \text{for } \mu \in S \quad \text{and } n = 0, 1, 2, \dots$$

Proof. Because of Theorem 7.1 it is sufficient to prove that

$$\|(C(\mu) - zI)^{-1}\| \leq \frac{C}{|z| - 1} \quad \text{for } \mu \in S \quad \text{and } |z| > 1.$$

To show this, we make use of the inequality (Kato (1960), see Exercise 2)

$$\|(C(\mu) - zI)^{-1}\| \leq \frac{(\|C(\mu)\| + |z|)^{k-1}}{|\det(C(\mu) - zI)|}.$$

$\|C(\mu)\|$ is uniformly bounded for $\mu \in S$. Therefore it suffices to show that

$$\varphi(\mu) = \inf_{|z| > 1} \frac{|\det(C(\mu) - zI)|}{|z|^{k-1}(|z| - 1)} \quad (7.21)$$

is bounded away from zero for all $\mu \in S$. For $|z| \rightarrow \infty$ the expression in (7.21) tends to 1 and so poses no problem. Further, observe that

$$|\det(C(\mu) - zI)| = \left| \prod_{j=1}^k (z - \zeta_j(\mu)) \right|, \quad (7.22)$$

where $\zeta_j(\mu)$ are the eigenvalues of $C(\mu)$, i.e., the roots of

$$\sum_{i=0}^k (\alpha_i - \mu\beta_i)\zeta^i = 0. \quad (7.23)$$

By definition of the stability region S , the values $\zeta_j(\mu)$ lie, for $\mu \in S$, inside the closed unit disc and those on the unit circle are well separated. Therefore, for fixed $\mu_0 \in S$, only one of the $\zeta_j(\mu_0)$ can be close to a z with $|z| > 1$. The corresponding factor in (7.22) will be minorized by $|z| - 1$, the other factors are bounded away from zero. By continuity of the $\zeta_j(\mu)$, the same holds for all $\mu \in S$ in a sufficiently small neighbourhood $V(\mu_0)$ of μ_0 . Hence $\varphi(\mu) \geq a > 0$ for $\mu \in V(\mu_0) \cap S$. Since S is closed (compact in $\overline{\mathbb{C}}$) it is covered by a finite number of $V(\mu_0)$. Consequently $\varphi(\mu) \geq a > 0$ for all $\mu \in S$, which completes the proof of the theorem. \square

Remark. The hypothesis “ S is closed in $\overline{\mathbb{C}}$ ” is usually satisfied. For methods which do *not* satisfy this hypothesis (see e.g., Exercise 2 of Sect. V.1 or Dahlquist, Mingyou & LeVeque (1981)) the above lemma remains valid on closed subsets $D \subset S \subset \overline{\mathbb{C}}$.

The estimate of this lemma can be improved, if we consider closed sets D lying in the interior of S .

Lemma 7.4. *Let S be the stability region of a method (ϱ, σ) . If $D \subset \text{Int } S$ is closed in $\overline{\mathbb{C}}$, then there exist constants M and κ ($0 < \kappa < 1$) such that*

$$\|C(\mu)^n\| \leq M\kappa^n \quad \text{for } \mu \in D \quad \text{and} \quad n = 0, 1, 2, \dots$$

Proof. If μ lies in the interior of S , all roots of (7.23) satisfy $|\zeta_j(\mu)| < 1$ (maximum principle). Since D is closed, this implies the existence of $\varepsilon > 0$ such that

$$D \subset S_\varepsilon = \{\mu \in \overline{\mathbb{C}}; |\zeta_j(\mu)| \leq 1 - 2\varepsilon, j = 1, \dots, k\}.$$

We now consider $R(\mu) = \kappa^{-1}C(\mu)$ with $\kappa = 1 - \varepsilon$. The eigenvalues of $R(\mu)$ satisfy $|\kappa^{-1}\zeta_j(\mu)| \leq (1 - 2\varepsilon)/(1 - \varepsilon) < 1 - \varepsilon$ for $\mu \in S_\varepsilon$. As in the proof of Lemma 7.3 (more easily, because $R(\mu)$ has no eigenvalues of modulus 1) we conclude that $R(\mu)$ is uniformly power bounded for $\mu \in S_\varepsilon$. This implies the statement. \square

Since the origin is never in the interior of S , we add the following estimate for its neighbourhood:

Lemma 7.5. *Suppose that the method (ϱ, σ) is consistent and strictly stable (see Sect. III.9, Assumption A1). Then there exists a neighbourhood V of 0 and constants M and a such that*

$$\|C(\mu)^n\| \leq Me^{n(\text{Re } \mu + a|\mu|^2)} \quad \text{for } \mu \in V \quad \text{and} \quad n = 0, 1, 2, \dots$$

Proof. Since the method is strictly stable there exists a compact neighbourhood V of 0, in which $|\zeta_j(\mu)| < |\zeta_1(\mu)|$ for $j = 2, \dots, k$ ($\zeta_j(\mu)$ are the roots of (7.23)). The matrix $R(\mu) = \zeta_1(\mu)^{-1}C(\mu)$ then has a simple eigenvalue 1 and all other eigenvalues are strictly smaller than 1. As in the proof of Lemma 7.3 we obtain $\|R(\mu)^n\| \leq M$ and consequently $\|C(\mu)^n\| \leq M|\zeta_1(\mu)|^n$ for $\mu \in V$. The stated estimate now follows from $\zeta_1(\mu) = e^\mu + O(\mu^2)$. \square

Global Error for Prothero and Robinson Problem

The above lemmas permit us to continue our analysis of Eq. (7.16). Whenever we consider λ and h such that their product λh varies in a closed subset of S , it follows that

$$\|E_{m+1}\| \leq M \left(\|E_0\| + \sum_{j=0}^m \|\Delta_j\| \right) \quad (7.24)$$

(Lemma 7.3). If $h\lambda$ varies in a closed subset of the interior of S , we have the better estimate

$$\|E_{m+1}\| \leq M \left(\kappa^{m+1} \|E_0\| + \sum_{j=0}^m \kappa^{m-j} \|\Delta_j\| \right) \quad \text{with some } \kappa < 1 \quad (7.25)$$

(Lemma 7.4). The resulting asymptotic estimates for the global errors $e_m = y_m - y(x_m)$ for $mh \leq \text{Const}$ are presented in Table 7.1 (p denotes the classical order, p_D the differentiation order and p_I the interpolation order of Sect. V.6). We assume that the initial values are exact and that simultaneously $h\lambda \rightarrow \infty$ and $h \rightarrow 0$. This is the most interesting situation because any reasonable method for stiff problems should integrate the equation with step sizes h such that $h\lambda$ is large. We distinguish two cases:

- (A) the half-ray $\{h\lambda; h > 0, |h\lambda| \geq c\} \cup \{\infty\}$ lies in S (Lemma 7.3 is applicable, i.e., Eq. (7.24)).
- (B) ∞ is an interior point of S (estimate (7.25) is applicable; the global error $\|E_m\|$ is essentially equal to the last term in the sum of (7.25)).

Table 7.1. Error for (7.1) when $h\lambda \rightarrow \infty$ and $h \rightarrow 0$

Method	error	(A)	(B)
multistep	e_m	$\mathcal{O}(\lambda ^{-1} h^{p-1})$	$\mathcal{O}(\lambda ^{-1} h^p)$
one-leg	e_m	$\mathcal{O}(h^{p_I+1} + \lambda ^{-1} h^{p_D-1})$	$\mathcal{O}(h^{p_I+1} + \lambda ^{-1} h^{p_D})$

We remark that the global error of the multistep method contains a factor $|\lambda|^{-1}$, so that the error decreases if $|\lambda|$ increases (“the stiffer the better”). The estimate in case (A) for one-leg methods is obtained by the use of Recursion (7.10).

Convergence for Linear Systems with Constant Coefficients

The extension of the above results to linear systems

$$y' = Ay + g(x), \quad y(x_0) = y_0 \quad (7.26)$$

is straightforward, if we assume that the matrix A is diagonalizable. The following results have been derived by Crouzeix & Raviart (1980).

Theorem 7.6. *Suppose that the multistep method (ϱ, σ) is of order p , $A(\alpha)$ -stable and stable at infinity. If the matrix A of (7.26) is diagonalizable (i.e., $T^{-1}AT = \text{diag}(\lambda_1, \dots, \lambda_n)$) with eigenvalues satisfying*

$$|\arg(-\lambda_i)| \leq \alpha \quad \text{for } i = 1, \dots, n,$$

then there exists a constant M (depending only on the method) such that for all $h > 0$ the global error satisfies

$$\|y(x_m) - y_m\| \leq M \cdot \|T\| \cdot \|T^{-1}\| \left(\max_{0 \leq j < k} \|y(x_j) - y_j\| + h^p \int_{x_0}^{x_m} \|y^{(p+1)}(\xi)\| d\xi \right).$$

Proof. The transformation $y = Tz$ decouples the system (7.26) into n scalar equations

$$z'_i = \lambda_i z_i + (T^{-1}g)_i(x). \quad (7.27)$$

Since this transformation leaves the numerical solution invariant, it suffices to consider Eq. (7.27). Lemma 7.3 yields the power boundedness

$$\|C(h\lambda_i)^m\| \leq M_0 \quad \text{for } h > 0, \quad i = 1, \dots, n \quad \text{and} \quad m \geq 0. \quad (7.28)$$

The discretization error $\delta_{LM}(x)$ (Eq. (7.5)) can be written as

$$\delta_{LM}(x) = h^{p+1} \int_0^k K_p(s) z_i^{(p+1)}(x + sh) ds, \quad (7.29)$$

where $K_p(s)$ is the Peano-kernel of the multistep method (Theorem III.2.8). By $A(\alpha)$ -stability we have $\alpha_k \cdot \beta_k > 0$, so that $|\alpha_k - h\lambda_i\beta_k|^{-1} \leq |\alpha_k|^{-1}$. This together with (7.29) implies that

$$\|\Delta_j\| \leq Ch^p \int_{x_j}^{x_{j+k}} |z_i^{(p+1)}(\xi)| d\xi, \quad (7.30)$$

where C depends only on the method. The estimates (7.28) and (7.30) inserted into (7.16) yield a bound for the global error of (7.27), which, by backsubstitution into the original variables, proves the statement. \square

Because of its exponentially decaying term, the following estimate is especially useful in the case when large time intervals are considered (or when the starting values do not lie on the exact solution).

Theorem 7.7. *Let the multistep method (ϱ, σ) be of order $p \geq 1$, $A(\alpha)$ -stable and strictly stable at zero and at infinity (i.e., $\sigma(\zeta) = 0$ implies $|\zeta| < 1$). If the matrix A of (7.26) is diagonalizable ($T^{-1}AT = \text{diag}(\lambda_1, \dots, \lambda_n)$) with eigenvalues λ_i satisfying*

$$|\arg(-\lambda_i)| \leq \gamma < \alpha, \quad \text{Re } \lambda_i \leq -\hat{\lambda} < 0$$

then, for given $h_0 > 0$, there exist constants M and $\nu > 0$ such that for $0 < h \leq h_0$

$$\begin{aligned} \|y(x_m) - y_m\| \leq & M \cdot \|T\| \cdot \|T^{-1}\| \cdot \left(e^{-\nu(x_m - x_0)} \cdot \max_{0 \leq j < k} \|y(x_j) - y_j\| \right. \\ & \left. + h^p \int_{x_0}^{x_m} e^{-\nu(x_m - \xi)} \|y^{(p+1)}(\xi)\| d\xi \right). \end{aligned}$$

Remark. The constants M and ν may depend on $\gamma, \hat{\lambda}, h_0$ and on the method, but they are independent of the eigenvalues λ_i and of the length $x_m - x_0$ of the integration interval.

Proof. By Lemma 7.5 there exists an $r > 0$ such that

$$\|C(h\lambda_i)^m\| \leq M_0 e^{-mh\hat{\lambda}/2} \quad \text{for } |h\lambda_i| \leq r \quad (7.31)$$

(observe that $|\mu| \leq \text{Const} \cdot |\text{Re } \mu|$, if $|\arg(-\mu)| \leq \gamma < \pi/2$). Since

$$D = \{\mu; |\arg(-\mu)| \leq \gamma, |\mu| \geq r\} \cup \{\infty\}$$

lies in the interior of the stability region S , it follows from Lemma 7.4 that

$$\|C(h\lambda_i)^m\| \leq M_1 \varrho^m \quad \text{for } |h\lambda_i| \geq r \quad (7.32)$$

with some $\varrho < 1$. Combining the estimates (7.31) and (7.32) we get

$$\|C(h\lambda_i)^m\| \leq M e^{-mh\nu} \quad \text{for } 0 < h \leq h_0, \quad (7.33)$$

where $M = \max(M_0, M_1)$ and $\nu = \min(\hat{\lambda}/2, -\ln \varrho/h_0)$. Using (7.33) instead of (7.28) and $mh = x_m - x_0$, the statement now follows as in the proof of Theorem 7.6. \square

Matrix Valued Theorem of von Neumann

An interesting contractivity result is obtained by the following matrix valued version of a theorem of von Neumann (Theorem IV.11.2).

We consider the Euclidean scalar product $\langle \cdot, \cdot \rangle$ on \mathbb{R}^n , the norm $\|\cdot\|_G$ on \mathbb{R}^k which is defined by a symmetric, positiv definite matrix G , and

$$\|u\|_G = \sqrt{\sum_{i,j=1}^k g_{ij} \langle u_i, u_j \rangle} \quad \text{for } u = (u_1, \dots, u_k)^T \in \mathbb{R}^{nk}. \quad (7.34)$$

The corresponding operator norms are denoted by the same symbols.

Theorem 7.8 (O. Nevanlinna 1985). *Let $C(\mu) = (c_{ij}(\mu))_{i,j=1}^k$ be a matrix whose elements are rational functions of μ . If*

$$\|C(\mu)\|_G \leq 1 \quad \text{for} \quad \operatorname{Re} \mu \leq 0, \quad (7.35)$$

then

$$\|C(A)\|_G \leq 1 \quad (7.36)$$

for all matrices A such that

$$\operatorname{Re} \langle y, Ay \rangle \leq 0 \quad \text{for} \quad y \in \mathbb{C}^n. \quad (7.37)$$

Remark. If $C(\mu)$ is the companion matrix of a G -stable method (ϱ, σ) , the result follows from Theorem 6.7 and Exercise 3 below (“It would be interesting to have a more operator-theoretical proof of this.” Dahlquist & Söderlind 1982).

Proof. This is a straight-forward extension of Crouzeix’s proof of Theorem IV.11.2. We first suppose that A is normal, so that $A = QDQ^*$ with a unitary matrix Q and a diagonal matrix $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. In this case we have

$$\|C(A)\|_G = \|(I \otimes Q)C(D)(I \otimes Q^*)\|_G = \|C(D)\|_G. \quad (7.38)$$

With the permutation matrix $P = (I \otimes e_1, \dots, I \otimes e_n)$ (where I is the k -dimensional identity and e_j is the n -dimensional j -th unit vector) the matrix $C(D)$ is transformed to block-diagonal form according to

$$P^*C(D)P = \operatorname{blockdiag}(C(\lambda_1), \dots, C(\lambda_n)).$$

We further have $P^*(G \otimes I)P = I \otimes G$. This implies that

$$P^*C(D)^*(G \otimes I)C(D)P = \operatorname{blockdiag}(C(\lambda_1)^*GC(\lambda_1), \dots)$$

and hence also

$$\|C(D)\|_G = \max_{i=1, \dots, n} \|C(\lambda_i)\|_G. \quad (7.39)$$

The statement now follows from (7.38) and (7.39), because $\operatorname{Re} \lambda_i \leq 0$ by (7.37).

For a general A we consider $A(\omega) = \frac{\omega}{2}(A + A^*) + \frac{1}{2}(A - A^*)$ and define the rational function

$$\varphi(\omega) = \langle u, C(A(\omega))v \rangle_G = u^*(G \otimes I)C(A(\omega))v.$$

The statement of the theorem can then be deduced exactly as in the proof of Theorem IV.11.2. \square

This theorem can be used to derive convergence results for differential equations (7.26) with A satisfying (7.37). Indeed, if the method (ϱ, σ) is A -stable, the companion matrix (7.13) satisfies $\|C(\mu)\|_G \leq 1$ for $\operatorname{Re} \mu \leq 0$ in some suitable

norm (Exercise 3). The above theorem then implies $\|C(hA)\|_G \leq 1$ and Formula (7.16) with λ replaced by A yields the estimate

$$\|E_{m+1}\|_G \leq \|E_0\|_G + \sum_{j=0}^m \|\Delta_j\|_G. \quad (7.40)$$

This proves convergence, because Δ_j can be bounded as in (7.30).

Discrete Variation of Constants Formula

A second approach to convergence results of linear multistep methods is by the use of a discrete variation of constants formula. This is an extension of the classical proofs for nonstiff problems (Dahlquist 1956, Henrici 1962) to the case $\mu \neq 0$. It has been developed by Crouzeix & Raviart (1976), and more recently by Lubich (1988, 1991).

We consider the error equation (cf. (7.13))

$$\sum_{i=0}^k (\alpha_i - \mu\beta_i) e_{m+i} = d_{m+k} \quad \text{for } m \geq 0, \quad (7.41)$$

and extend this relation to negative m by putting $e_j = 0$ (for $j < 0$) and by defining d_0, \dots, d_{k-1} according to (7.41). The main idea is now to introduce the generating power series

$$e(\zeta) = \sum_{j \geq 0} e_j \zeta^j, \quad d(\zeta) = \sum_{j \geq 0} d_j \zeta^j$$

so that (7.41) becomes the m -th coefficient of the identity

$$(\varrho(\zeta^{-1}) - \mu\sigma(\zeta^{-1}))e(\zeta) = \zeta^{-k}d(\zeta). \quad (7.42)$$

This gives

$$e(\zeta) = (\varrho(\zeta^{-1}) - \mu\sigma(\zeta^{-1}))^{-1} \zeta^{-k} d(\zeta) = r(\zeta, \mu) d(\zeta) \quad (7.43)$$

and allows to compute easily e_m in terms of d_j as

$$e_m = \sum_{j=0}^m r_{m-j}(\mu) d_j. \quad (7.43')$$

Here $r_j(\mu)$ are the coefficients of the *discrete resolvent*

$$r(\zeta, \mu) = (\delta(\zeta) - \mu)^{-1} \frac{\zeta^{-k}}{\sigma(\zeta^{-1})} = \sum_{j \geq 0} r_j(\mu) \zeta^j, \quad (7.44)$$

where

$$\delta(\zeta) = \frac{\varrho(\zeta^{-1})}{\sigma(\zeta^{-1})} = \frac{\alpha_0 \zeta^k + \dots + \alpha_{k-1} \zeta + \alpha_k}{\beta_0 \zeta^k + \dots + \beta_{k-1} \zeta + \beta_k}. \quad (7.45)$$

Since $(\varrho(\zeta^{-1}) - \mu\sigma(\zeta^{-1}))r_j(\zeta, \mu) = \zeta^{-k}$, the coefficients $r_j(\mu)$ can be interpreted as the numerical solution y_j of the multistep method applied to the homogeneous equation $y' = \mu y$ with step size $h = 1$, and with starting values $y_{-k+1} = \dots = y_{-1} = 0$, $y_0 = (\alpha_k - \mu\beta_k)^{-1}$.

Formula (7.43') can be used to estimate e_m , if appropriate bounds for the coefficients $r_j(\mu)$ of the discrete resolvent are known. Such bounds are given in the following lemma.

Lemma 7.9. *Let $S \subset \overline{\mathbb{C}}$ denote the stability region of the multistep method.*

a) *If S is closed in $\overline{\mathbb{C}}$ then*

$$|r_j(\mu)| \leq \frac{M}{1 + |\mu|} \quad \text{for } \mu \in S \quad \text{and } j = 0, 1, 2, \dots$$

b) *If $D \subset \text{Int } S$ is closed in $\overline{\mathbb{C}}$ then there exists a constant κ ($0 < \kappa < 1$) such that*

$$|r_j(\mu)| \leq \frac{M\kappa^j}{1 + |\mu|} \quad \text{for } \mu \in D \quad \text{and } j = 0, 1, 2, \dots$$

c) *If the method is strictly stable then there exists a neighbourhood V of 0 such that*

$$|r_j(\mu)| \leq M e^{j(\text{Re } \mu + a|\mu|^2)} \quad \text{for } \mu \in V \quad \text{and } j = 0, 1, 2, \dots$$

The constants M , κ , and a are independent of j and μ .

Proof. The estimates for $|r_j(\mu)|$ in (a), (b), and (c) can easily be deduced from Lemmas 7.3, 7.4, and 7.5 because $r_j(\mu)$ is the numerical solution for the problem $y' = \mu y$ with step size $h = 1$ and starting values $y_{-k+1} = \dots = y_{-1} = 0$, $y_0 = (\alpha_k - \mu\beta_k)^{-1}$.

As noted by Crouzeix & Raviart (1976) and Lubich (1988) the estimates of Lemma 7.9 can be proved *directly*, without any use of the Kreiss matrix theorem. We illustrate these ideas by proving statement (b) (for a proof of statement (a) see Exercise 4).

By definition of the stability region the function $\zeta^k(\varrho(\zeta^{-1}) - \mu\sigma(\zeta^{-1}))$ does not vanish for $|\zeta| \leq 1$ if $\mu \in \text{Int } S$. Therefore there exists a κ ($0 < \kappa < 1$) such that $\zeta^k(\varrho(\zeta^{-1}) - \mu\sigma(\zeta^{-1}))$ has no zeros in the disk $|\zeta| \leq 1/\kappa$. Hence, for $\mu \in D$

$$\sup_{|\zeta| \leq 1/\kappa} |(\varrho(\zeta^{-1}) - \mu\sigma(\zeta^{-1}))^{-1} \zeta^{-k}| \leq \frac{M}{1 + |\mu|},$$

and Cauchy's integral formula

$$r_j(\mu) = \frac{1}{2\pi i} \int_{|\zeta|=1/\kappa} (\varrho(\zeta^{-1}) - \mu\sigma(\zeta^{-1}))^{-1} \zeta^{-k} \zeta^{-j-1} d\zeta \quad (7.46)$$

yields the desired estimate. \square

The use of the discrete resolvent allows elegant convergence proofs for linear multistep methods. We shall demonstrate this for the linear problem (7.26) where the matrix A satisfies

$$\|(sI - A)^{-1}\| \leq \frac{M}{1 + |s|} \quad \text{for } |\arg(s - c)| \leq \pi - \alpha' \quad (7.47)$$

with some $c \in \mathbb{R}$. This is a common assumption in the theory of holomorphic semigroups for parabolic problems (see e.g., Kato (1966) or Pazy (1983)). If all eigenvalues λ_i of A satisfy $|\arg(\lambda_i - c) - \pi| < \alpha'$ then Condition (7.47) is satisfied with a constant M depending on the matrix A (Exercise 2). The following theorem, which was communicated to us by Ch. Lubich, is an improvement of results of Crouzeix & Raviart (1976).

Theorem 7.10. *Let the multistep method be of order $p \geq 1$, $A(\alpha)$ -stable and strictly stable at zero and at infinity. If the matrix A of (7.26) satisfies (7.47) with $\alpha' < \alpha$, then there exist constants C , h_0 , and γ (γ of the same sign as c in (7.47)), which depend only on M , c , α' and the method, such that for $h \leq h_0$ the global error satisfies*

$$\begin{aligned} & \|y(x_m) - y_m\| \\ & \leq C(e^{\gamma x_m} \max_{0 \leq j < k} \|y(x_j) - y_j\| + h^p \int_{x_0}^{x_m} e^{\gamma(x_m - \xi)} \|y^{(p+1)}(\xi)\| d\xi). \end{aligned}$$

Moreover, if $c \leq 0$, then h_0 can be chosen arbitrarily.

Proof. The global error $e_m = y(x_m) - y_m$ satisfies

$$\sum_{i=0}^k (\alpha_i - hA\beta_i) e_{m+i} = d_{m+k}$$

where

$$\|d_{m+k}\| \leq Ch^p \int_{x_m}^{x_{m+k}} \|y^{(p+1)}(\xi)\| d\xi, \quad m \geq 0 \quad (7.48)$$

and d_0, \dots, d_{k-1} are linear combinations of the e_j and hAe_j with $j < k$. We split these expressions into

$$d_\ell = d'_\ell + hAd''_\ell \quad \text{for } \ell < k,$$

so that d'_ℓ and d''_ℓ are linear combinations of the e_j ($j < k$) only. We also put $d'_\ell = d_\ell$ and $d''_\ell = 0$ for $\ell \geq k$. The analysis at the beginning of this subsection (Eq. (7.43)) then shows that

$$e(\zeta) = r(\zeta, hA)d'(\zeta) + r(\zeta, hA)hAd''(\zeta), \quad (7.49)$$

where as in the scalar case

$$r(\zeta, hA) = (\delta(\zeta)I - hA)^{-1} \frac{\zeta^{-k}}{\sigma(\zeta^{-1})} = \sum_{j \geq 0} r_j(hA) \zeta^j. \quad (7.50)$$

We now apply Lemma 7.11 below with $\Phi(s) = (sI - A)^{-1}$. By assumption the estimate (7.57) holds with $\beta = 1$ so that

$$\|r_j(hA)\| \leq Ce^{\gamma j h}. \quad (7.51)$$

The second term in (7.49) can be written as

$$r(\zeta, hA)hA(\delta(\zeta))^{-1}\delta(\zeta)d''(\zeta) = r'(\zeta, hA)\widehat{d}(\zeta) \quad (7.52)$$

where

$$\begin{aligned} r'(\zeta, hA) &= (\delta(\zeta)I - hA)^{-1}hA(\delta(\zeta))^{-1}\frac{\zeta^{-k}}{\sigma(\zeta^{-1})} = \sum_{j \geq 0} r'_j(hA)\zeta^j \\ \widehat{d}(\zeta) &= \delta(\zeta)d''(\zeta) = \sum_{j \geq 0} \widehat{d}_j\zeta^j. \end{aligned} \quad (7.53)$$

We apply Lemma 7.11 again, this time to

$$\Phi(s) = (sI - A)^{-1}As^{-1} = (sI - A)^{-1} - s^{-1}I.$$

Condition (7.57) is satisfied with $\beta = 1$ so that

$$\|r'_j(hA)\| \leq C'e^{\gamma j h}. \quad (7.54)$$

The coefficients δ_j of $\delta(\zeta)$ are exponentially decaying because all zeros of $\sigma(\zeta)$ lie in $|\zeta| < 1$. Consequently, we have

$$\|\widehat{d}_j\| \leq \kappa^j \widehat{C} \max_{0 \leq \ell < k} \|e_\ell\| \quad (7.55)$$

with some $\kappa < 1$. The coefficient of ζ^m in (7.49) gives

$$e_m = \sum_{j=0}^m r_{m-j}(hA)d'_j + \sum_{j=0}^m r'_{m-j}(hA)\widehat{d}_j.$$

Inserting the estimates (7.48), (7.51), (7.54), and (7.55) proves the statement. \square

We still have to prove the estimates for $r_j(hA)$ and $r'_j(hA)$. For this we let $\Phi(s)$ be some analytic (scalar-, vector-, or matrix-valued) function and consider the coefficients of

$$\Phi(\delta(\zeta)/h) \cdot \frac{\zeta^{-k}}{\sigma(\zeta^{-1})} = h \sum_{j \geq 0} \varphi_j(h)\zeta^j. \quad (7.56)$$

We then have the following result.

Lemma 7.11 (Lubich 1991). *Assume that the multistep method is $A(\alpha)$ -stable and strictly stable at zero and at infinity. Further suppose that $\Phi(s)$ is analytic in a sector $|\arg(s - c)| < \pi - \alpha'$ with $\alpha' < \alpha$, $c \in \mathbb{R}$ and there satisfies*

$$\|\Phi(s)\| \leq M \cdot |s|^{-\beta} \quad \text{for some } \beta > 0. \quad (7.57)$$

Then the coefficients $\varphi_j(h)$ of (7.56) are bounded for $h \leq h_0$ (sufficiently small) by

$$\|\varphi_j(h)\| \leq C \cdot (jh)^{\beta-1} e^{\gamma jh} \quad \text{for } j \geq 1, \quad (7.58)$$

and for $j = 0$ the same bound holds as for $j = 1$. The constants C , γ , and h_0 depend only on α' , c , M , β , and the multistep method. Moreover, if $c < 0$, then also $\gamma < 0$, and the result holds for arbitrary h_0 .

Proof. By $A(\alpha)$ -stability we have $\beta_k/\alpha_k > 0$, so that $\delta(0)/h$ lies in the region of analyticity of Φ for $h \leq h_0$. Cauchy's integral formula thus gives

$$\Phi(\delta(\zeta)/h) = \frac{1}{2\pi i} \int_{\Gamma} (\delta(\zeta)/h - \lambda)^{-1} \Phi(\lambda) d\lambda \quad (7.59)$$

where Γ is a suitable contour from " $\infty \cdot e^{-i(\pi-\alpha')}$ " to " $\infty \cdot e^{i(\pi-\alpha')}$ " within the sector of analyticity of Φ and does not meet the origin (see Fig. 7.1; observe that $\Phi(s)$ decays sufficiently rapidly at infinity). Multiplying (7.59) by $\zeta^{-k}/\sigma(\zeta^{-1})$ and comparing coefficients of equal powers of ζ yields the representation

$$\varphi_j(h) = \frac{1}{2\pi i} \int_{\Gamma} r_j(h\lambda) \Phi(\lambda) d\lambda, \quad j \geq 0, \quad (7.60)$$

which is a discrete analogue of the Laplace inversion formula. We next substitute $\omega = jh\lambda$ (for $j = 0$ we put $\omega = h\lambda$) so that with $\Gamma_j = jh \cdot \Gamma$ Eq. (7.60) becomes

$$\varphi_j(h) = \frac{1}{2\pi i} \int_{\Gamma_j} r_j\left(\frac{\omega}{j}\right) \Phi\left(\frac{\omega}{jh}\right) \frac{d\omega}{jh}, \quad j \geq 1, \quad (7.61)$$

and the use of (7.57) yields

$$\|\varphi_j(h)\| \leq \frac{M}{2\pi} (jh)^{\beta-1} \int_{\Gamma_j} \left| r_j\left(\frac{\omega}{j}\right) \right| \cdot |\omega|^{-\beta} \cdot |d\omega|. \quad (7.62)$$

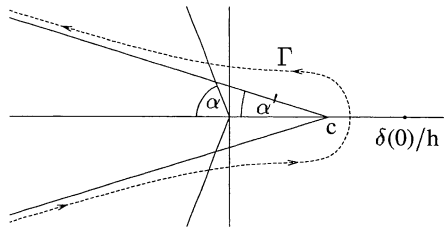


Fig. 7.1. Contour Γ in Formula (7.59)

We still have to show that the integral in (7.62) is bounded by $C \cdot e^{\gamma jh}$. For this we split it into two parts: the first one corresponds to those ω such that ω/j lies in a closed subset of the interior of the stability domain of the method. There we can use Lemma 7.9b so that the corresponding part of the integral in (7.62) is bounded by

$$j \cdot \kappa^j \int |\omega|^{-\beta-1} |d\omega| \leq C e^{\gamma jh} \quad \text{for } h \leq h_0.$$

For the remaining part, the argument $\omega/j = h\lambda$ of r_j in (7.62) lies, for sufficiently small h_0 , in a neighbourhood V of the origin, where the estimate of Lemma 7.9c holds. For $jh \geq 1$ we thus obtain the bound

$$\int e^{\operatorname{Re} \omega + a|\omega|^2/j} |\omega|^{-\beta} |d\omega| \leq C e^{\gamma jh},$$

because $\operatorname{Re} \omega = jh \operatorname{Re} \lambda$, $|\omega|^2/j \leq jh \cdot \text{Const}$ and $|\omega| \geq |\lambda|$ is bounded away from the origin. For small jh the contour Γ_j comes arbitrarily close to the origin so that a more refined estimate is required. The idea is to replace the corresponding part of Γ_j (in (7.61) and hence also in (7.62)) by an equivalent contour which is independent of $jh \in [h, 1]$, has a positive distance to the origin and remains in the neighbourhood V . The corresponding integral is thus bounded by some constant. \square

Remark 7.12. In Lemma 7.11 it is sufficient to require the analyticity of $\Phi(s)$ and the estimate (7.57) in a sector $|\arg(s-c)| < \pi - \alpha'$, where some compact neighbourhood of the origin is removed. We just have to take the contour Γ in (7.59) so that it lies outside this compact neighbourhood of 0. In this situation, the constant γ may be positive also if $c < 0$.

Exercises

1. Prove the Cauchy integral formula (7.18) in the case where all eigenvalues λ of A satisfy $|\lambda| \leq 1$ and the contour of integration is the circle $|z| = \varrho$ with $\varrho > 1$.

Hint. Integrate the identity

$$z^n (zI - A)^{-1} = \sum_{j=0}^{\infty} A^j z^{n-j-1}.$$

2. (Kato 1960). For a non-singular $k \times k$ -matrix B show that in the Euclidean norm

$$\|B^{-1}\| \leq \frac{\|B\|^{k-1}}{|\det B|}.$$

Hint. Use the singular value decomposition of B , i.e., $B = U^T \Lambda V$, where U and V are orthogonal and $\Lambda = \operatorname{diag}(\sigma_1, \dots, \sigma_k)$ with $\sigma_1 \geq \sigma_2 \geq \dots > \sigma_k > 0$.

3. A method (ϱ, σ) is called *A-contractive* in the norm $\|\cdot\|_G$ (Nevanlinna & Liniger 1978-79, Dahlquist & Söderlind 1982), if

$$\|C(\mu)\|_G \leq 1 \quad \text{for} \quad \operatorname{Re} \mu \leq 0$$

where $C(\mu)$ is the companion matrix (7.13).

- a) Prove that a method (ϱ, σ) is A -contractive for some positive definite matrix G , if and only if it is A -stable.
 b) Compute the contractivity region

$$\{\mu \in \mathbb{C}; \|C(\mu)\|_G \leq 1\}$$

for the 2-step BDF method with G given in (6.20). Observe that it is strictly smaller than the stability domain.

Result. The contractivity region is $\{\mu \in \mathbb{C}; \operatorname{Re} \mu \leq 0\}$.

4. Give a direct proof for the statement of Lemma 7.9a.

Hint. Observe that

$$r(\zeta, \mu) = \frac{1}{\alpha_k - \mu\beta_k} \prod_{i=1}^k \frac{1}{(1 - \zeta \cdot \zeta_i(\mu))}, \quad (7.63)$$

where $\zeta_1(\mu), \dots, \zeta_k(\mu)$ are the k zeros of $\varrho(\zeta) - \mu\sigma(\zeta)$. If $\mu_0 \in \operatorname{Int} S$ then there exists a neighbourhood \mathcal{U} of μ_0 such that $|\zeta_i(\mu)| \leq a < 1$ for all i and $\mu \in \mathcal{U}$. Hence the coefficients $r_j(\mu)$ are bounded. For $\mu_0 \in \partial S$ we have $|\zeta_i(\mu_0)| = 1$ for, say, $i = 1, \dots, \ell$ with $1 \leq \ell \leq k$. These ℓ zeros are simple for all μ in a sufficiently small neighbourhood \mathcal{U} of μ_0 and the other zeros satisfy $|\zeta_i(\mu)| \leq a < 1$ for $\mu \in \mathcal{U} \cap S$. A partial fraction decomposition

$$r(\zeta, \mu) = \frac{1}{\alpha_k - \mu\beta_k} \left(\sum_{i=1}^{\ell} \frac{c_i(\mu)}{1 - \zeta \cdot \zeta_i(\mu)} + s(\zeta, \mu) \right)$$

shows that

$$r_j(\mu) = \frac{1}{\alpha_k - \mu\beta_k} \left(\sum_{i=1}^{\ell} c_i(\mu) (\zeta_i(\mu))^j + s_j(\mu) \right), \quad (7.64)$$

where $s_j(\mu)$ are the coefficients of $s(\zeta, \mu)$. Since the function $s(\zeta, \mu)$ is uniformly bounded for $|\zeta| \leq 1$ and $\mu \in \mathcal{U} \cap S$, it follows from Cauchy's integral formula with integration along $|\zeta| = 1$ that $s_j(\mu)$ is bounded. The statement thus follows from (7.64) and the fact that a finite set of the family $\{\mathcal{U}\}_{\mu_0 \in S}$ covers S (Heine-Borel).

V.8 Convergence for Nonlinear Problems

In Sect. V.6 we have seen a convergence result for one-leg methods (Theorem 6.10) applied to nonlinear problems satisfying a one-sided Lipschitz condition. An extension to linear multistep methods has been given in Theorem 6.11. A different and direct proof of this result will be the first goal of this section. Unfortunately, such a result is valid only for A -stable methods (whose order cannot exceed two). The subsequent parts of this section are then devoted to convergence results for nonlinear problems, where the assumptions on the method are relaxed (e.g., $A(\alpha)$ -stability), but the class of problems considered is restricted. We shall present two different theories: the multiplier technique of Nevanlinna & Odeh (1981) and Lubich's perturbation approach via the discrete variation of constants formula (Lubich 1991).

Problems Satisfying a One-Sided Lipschitz Condition

Suppose that the differential equation $y' = f(x, y)$ satisfies

$$\operatorname{Re} \langle f(x, y) - f(x, z), y - z \rangle \leq \nu \|y - z\|^2 \quad (8.1)$$

for some inner product. We consider the linear multistep method

$$\sum_{i=0}^k \alpha_i y_{m+i} = h \sum_{i=0}^k \beta_i f(x_{m+i}, y_{m+i}) \quad (8.2)$$

together with its perturbed formula

$$\sum_{i=0}^k \alpha_i \hat{y}_{m+i} = h \sum_{i=0}^k \beta_i f(x_{m+i}, \hat{y}_{m+i}) + d_{m+k}. \quad (8.3)$$

The perturbations d_{m+k} can be interpreted as the influence of round-off, as the error due to the iterative solution of the nonlinear equation, or as the local discretization error (compare Eq. (7.5)). Taking the difference of (8.3) and (8.2) we obtain (for $m \geq 0$)

$$\sum_{i=0}^k \alpha_i \Delta y_{m+i} = h \sum_{i=0}^k \beta_i \Delta f_{m+i} + d_{m+k}, \quad (8.4)$$

where we have introduced the notation

$$\Delta y_j = \widehat{y}_j - y_j, \quad \Delta f_j = f(x_j, \widehat{y}_j) - f(x_j, y_j). \quad (8.5)$$

The one-sided Lipschitz condition cannot be used directly, because several Δf_j appear in (8.4) (in contrast to one-leg methods). In order to express *one* Δf_m in terms of Δy_j only we introduce the formal power series

$$\Delta y(\zeta) = \sum_{j \geq 0} \Delta y_j \zeta^j, \quad \Delta f(\zeta) = \sum_{j \geq 0} \Delta f_j \zeta^j, \quad d(\zeta) = \sum_{j \geq 0} d_j \zeta^j.$$

It is convenient to assume that $\Delta y_j = 0$, $\Delta f_j = 0$, $d_j = 0$ for negative indices and that d_0, \dots, d_{k-1} are defined by Eq. (8.4) with $m \in \{-k, \dots, -1\}$. Then Eq. (8.4) just compares the coefficient of ζ^m in the identity

$$\varrho(\zeta^{-1}) \Delta y(\zeta) = h\sigma(\zeta^{-1}) \Delta f(\zeta) + \zeta^{-k} d(\zeta). \quad (8.4')$$

Dividing (8.4') by $\sigma(\zeta^{-1})$ and comparing the coefficients of ζ^m yields

$$\sum_{j=0}^m \delta_{m-j} \Delta y_j = h\Delta f_m + \widetilde{d}_m, \quad (8.6)$$

where

$$\frac{\varrho(\zeta^{-1})}{\sigma(\zeta^{-1})} = \delta(\zeta) = \sum_{j \geq 0} \delta_j \zeta^j \quad (8.7)$$

as in (7.45) and

$$\frac{\zeta^{-k}}{\sigma(\zeta^{-1})} d(\zeta) = \widetilde{d}(\zeta) = \sum_{j \geq 0} \widetilde{d}_j \zeta^j. \quad (8.8)$$

In (8.6) Δf_m is now isolated as desired and we can take the scalar product of (8.6) with Δy_m . We then exploit the assumption (8.1) and obtain

$$\sum_{j=0}^m \delta_{m-j} \operatorname{Re} \langle \Delta y_j, \Delta y_m \rangle \leq h\nu \|\Delta y_m\|^2 + \operatorname{Re} \langle \widetilde{d}_m, \Delta y_m \rangle. \quad (8.9)$$

This allows us to prove the following estimate.

Lemma 8.1. *Let $\{\Delta y_j\}$ and $\{\Delta f_j\}$ satisfy (8.6) with δ_j given by (8.7). If*

$$\operatorname{Re} \langle \Delta f_m, \Delta y_m \rangle \leq \nu \|\Delta y_m\|^2, \quad m \geq 0,$$

and the method is A-stable, then there exist constants C and $C_0 > 0$ such that for $mh \leq x_{\text{end}} - x_0$ and $h\nu \leq C_0$,

$$\|\Delta y_m\| \leq C \sum_{j=0}^m \|\widetilde{d}_j\|.$$

Proof. We first reformulate the left-hand side of (8.9). For this we introduce $\{\Delta z_j\}$ by the relation

$$\sum_{i=0}^k \beta_i \Delta z_{m+i} = \Delta y_m, \quad m \geq 0 \quad (8.10)$$

and assume that $\Delta z_j = 0$ for $j < k$. With $\Delta z(\zeta) = \sum_j \Delta z_j \zeta^j$ this means that $\sigma(\zeta^{-1})\Delta z(\zeta) = \Delta y(\zeta)$. Consequently we also have

$$\delta(\zeta) \Delta y(\zeta) = \varrho(\zeta^{-1}) \Delta z(\zeta),$$

which is equivalent to

$$\sum_{j=0}^m \delta_{m-j} \Delta y_j = \sum_{i=0}^k \alpha_i \Delta z_{m+i}. \quad (8.11)$$

Inserting (8.11) and (8.10) into (8.9) yields

$$\begin{aligned} \operatorname{Re} \left\langle \sum_{i=0}^k \alpha_i \Delta z_{m+i}, \sum_{i=0}^k \beta_i \Delta z_{m+i} \right\rangle \\ \leq h\nu \left\| \sum_{i=0}^k \beta_i \Delta z_{m+i} \right\|^2 + \operatorname{Re} \left\langle \tilde{d}_m, \sum_{i=0}^k \beta_i \Delta z_{m+i} \right\rangle. \end{aligned}$$

By Theorem 6.7 the method (ϱ, σ) is G -stable, so that Eq. (6.21) can be applied. As in the proof of Lemma 6.9 this yields for $\Delta Z_m = (\Delta z_{m+k-1}, \dots, \Delta z_m)^T$ and $\nu \geq 0$

$$\|\Delta Z_{m+1}\|_G \leq (1 + C_1 h\nu) \|\Delta Z_m\|_G + C_2 \|\tilde{d}_m\|,$$

(if $\nu < 0$ replace ν by $\nu = 0$). But this implies

$$\|\Delta Z_{m+1}\|_G \leq C_3 \left(\|\Delta Z_0\|_G + \sum_{j=0}^m \|\tilde{d}_j\| \right).$$

By definition of Δz_j we have $\Delta Z_0 = 0$. The statement now follows from the fact that $\|\Delta y_m\| \leq C_4 (\|\Delta Z_{m+1}\|_G + \|\Delta Z_m\|_G)$. \square

This lemma allows a direct proof for the convergence of A -stable multistep methods which are strictly stable at infinity (compare Theorem 6.11).

Theorem 8.2. *Consider an A -stable multistep method of order p which is strictly stable at infinity. Suppose that the differential equation satisfies (8.1). Then there exists $C_0 > 0$ such that for $h\nu \leq C_0$*

$$\|y_m - y(x_m)\| \leq C \left(\max_{0 \leq j < k} \|y_j - y(x_j)\| + h \max_{0 \leq j < k} \|f(x_j, y_j) - y'(x_j)\| \right) + Mh^p.$$

The constant C depends on the method and, for $\nu > 0$, on the length $x_m - x_0$ of the integration interval; the constant M depends in addition on bounds for the $(p+1)$ -th derivative of the exact solution.

Proof. We put $\hat{y}_m = y(x_m)$ in (8.3). The perturbations thus become the local truncation errors $d_{m+k} = \delta_{LM}(x_m)$, where

$$\delta_{LM}(x) = \sum_{i=0}^k \alpha_i y(x + ih) - h \sum_{i=0}^k \beta_i y'(x + ih). \quad (8.12)$$

If the zeros of $\sigma(\zeta)$ all lie in the open unit disk, the coefficients of $\zeta^{-k}/\sigma(\zeta^{-1})$ are absolutely summable and by (8.8) we have

$$\sum_{j=0}^m \|\tilde{d}_j\| \leq C_1 \sum_{j=0}^m \|d_j\|.$$

The statement then follows from Lemma 8.1, from $\|\delta_{LM}(x)\| \leq Mh^{p+1}$, and from the fact that d_0, \dots, d_{k-1} are linear combinations of the $y_j - y(x_j)$ and $h(f(x_j, y_j) - y'(x_j))$ for $j < k$. \square

Multiplier Technique

... the best of all multipliers would be $\{1, -\eta\}$ with a very small $\eta > 0$; ... (Nevanlinna & Odeh 1981)

The above convergence proof is based on Eq. (8.6) and on the A -stability of the multistep method. How can we modify this proof in order to get convergence results also for methods which are not A -stable? This can be done by the so-called “multiplier technique”, introduced by Nevanlinna & Odeh (1981) and based on previous ideas of Nevanlinna (1977) and Odeh & Liniger (1977).

The main idea is the following: instead of multiplying scalarly the identity (8.6) by Δy_m , we multiply it by

$$\sum_{j=0}^m \mu_{m-j} \Delta y_j$$

where $\{\mu_j\}$ are the coefficients of a rational function (the multiplier)

$$\mu(\zeta) = \sum_{j \geq 0} \mu_j \zeta^j = \frac{\eta(\zeta^{-1})}{\tau(\zeta^{-1})} \quad (8.13)$$

(η and τ are polynomials). We obtain

$$\begin{aligned} \operatorname{Re} \left\langle \sum_{j=0}^m \delta_{m-j} \Delta y_j, \sum_{j=0}^m \mu_{m-j} \Delta y_j \right\rangle &= h \operatorname{Re} \left\langle \Delta f_m, \sum_{j=0}^m \mu_{m-j} \Delta y_j \right\rangle \\ &+ \left\langle \tilde{d}_m, \sum_{j=0}^m \mu_{m-j} \Delta y_j \right\rangle. \end{aligned} \quad (8.14)$$

Our next aim is to introduce new variables Δz_j such that the left-hand side of (8.14) becomes

$$\left\langle \sum_{j=0}^m \delta_{m-j} \Delta y_j, \sum_{j=0}^m \mu_{m-j} \Delta y_j \right\rangle = \left\langle \sum_{i=0}^{\ell} \tilde{\alpha}_i \Delta z_{m+i}, \sum_{i=0}^{\ell} \tilde{\beta}_i \Delta z_{m+i} \right\rangle. \quad (8.15)$$

Denoting

$$\tilde{\varrho}(\zeta) = \sum_{i=0}^{\ell} \tilde{\alpha}_i \zeta^i, \quad \tilde{\sigma}(\zeta) = \sum_{i=0}^{\ell} \tilde{\beta}_i \zeta^i, \quad (8.16)$$

the identity (8.15) certainly holds, if

$$\begin{aligned} \varrho(\zeta^{-1}) \Delta y(\zeta) &= \sigma(\zeta^{-1}) \tilde{\varrho}(\zeta^{-1}) \Delta z(\zeta) \\ \eta(\zeta^{-1}) \Delta y(\zeta) &= \tau(\zeta^{-1}) \tilde{\sigma}(\zeta^{-1}) \Delta z(\zeta). \end{aligned} \quad (8.17)$$

Dividing these two relations motivates the following definition of the new generating polynomials

$$\tilde{\varrho}(\zeta) = \varrho(\zeta) \tau(\zeta) / \chi(\zeta), \quad \tilde{\sigma}(\zeta) = \sigma(\zeta) \eta(\zeta) / \chi(\zeta). \quad (8.18)$$

Here $\chi(\zeta)$ denotes the greatest common divisor of $\varrho(\zeta) \tau(\zeta)$ and $\sigma(\zeta) \eta(\zeta)$. If we define $\Delta z_j = 0$ for $j < 0$ and the remaining Δz_j by

$$\chi(\zeta^{-1}) \Delta y(\zeta) = \sigma(\zeta^{-1}) \tau(\zeta^{-1}) \Delta z(\zeta) \quad (8.19)$$

the identity (8.15) holds for all m . Suppose now that the multistep method $(\tilde{\varrho}, \tilde{\sigma})$ is A -stable, then the left hand side of (8.14) can be minorized by the G -stability estimate (6.21) and we shall be able to derive convergence results. This motivates the following

Definition 8.3. The rational function $\mu(\zeta)$ of (8.13) is called a *multiplier* for (ϱ, σ) if $\mu(\zeta) \neq \varrho(\zeta^{-1})/\sigma(\zeta^{-1})$ and if the method $(\tilde{\varrho}, \tilde{\sigma})$, given by (8.18) is A -stable, i.e., if

$$\operatorname{Re} \left(\frac{1}{\mu(\zeta^{-1})} \cdot \frac{\varrho(\zeta)}{\sigma(\zeta)} \right) > 0 \quad \text{for } |\zeta| > 1. \quad (8.20)$$

A continuation of the above analysis yields the following convergence result.

Lemma 8.4. Let $\{\Delta y_j\}$ and $\{\Delta f_j\}$ satisfy (8.6) with δ_j given by (8.7). If

$$\sum_{m=0}^N \sum_{j=0}^m \mu_{m-j} \operatorname{Re} \langle \Delta f_m, \Delta y_j \rangle \leq 0 \quad \text{for all } N \geq 0$$

and if $\mu(\zeta)$ is a multiplier for the method, then there exists a constant C such that for $mh \leq x_{\text{end}} - x_0$

$$\|\Delta y_m\| \leq C \sum_{j=0}^m \|\tilde{d}_j\|.$$

Proof. Inserting (8.15) into (8.14) and using the estimate (6.21) for the A -stable method $(\tilde{\varrho}, \tilde{\sigma})$ yields for $\Delta Z_m = (\Delta z_{m+\ell-1}, \dots, \Delta z_m)^T$

$$\begin{aligned} \|\Delta Z_{m+1}\|_G^2 - \|\Delta Z_m\|_G^2 &\leq h \operatorname{Re} \left\langle \Delta f_m, \sum_{j=0}^m \mu_{m-j} \Delta y_j \right\rangle \\ &\quad + \|\tilde{d}_m\| \cdot \left\| \sum_{i=0}^{\ell} \tilde{\beta}_i \Delta z_{m+i} \right\|. \end{aligned} \quad (8.21)$$

Summing up this inequality from $m = 0$ to $m = N$ gives

$$\|\Delta Z_{N+1}\|_G^2 \leq C_1 \sum_{m=0}^N \|\tilde{d}_m\| \cdot (\|\Delta Z_{m+1}\|_G + \|\Delta Z_m\|_G),$$

because $\Delta Z_0 = 0$ by (8.19). This also implies

$$\max_{N \leq M} \|\Delta Z_{N+1}\|_G^2 \leq 2C_1 \sum_{m=0}^M \|\tilde{d}_m\| \cdot \max_{m \leq M} \|\Delta Z_{m+1}\|_G.$$

A division by $\max_{N \leq M} \|\Delta Z_{N+1}\|_G$ yields the desired estimate, because Δy_M is a linear combination of the elements of ΔZ_{M+1} . \square

The proof of Theorem 8.2 applied to the A -stable method $(\tilde{\varrho}, \tilde{\sigma})$ now yields:

Theorem 8.5 (Nevanlinna & Odeh 1981). *Consider a linear multistep method (8.2) of order p , which is strictly stable at infinity and has a multiplier $\mu(\zeta)$. Suppose that the differential equation satisfies*

$$\sum_{m=0}^N \sum_{j=0}^m \mu_{m-j} \operatorname{Re} \langle f(x_m, u_m) - f(x_m, v_m), u_j - v_j \rangle \leq 0 \quad (8.22)$$

for all $N \geq 0$ and for all sequences $\{u_j\}$ and $\{v_j\}$. Then we have

$$\|y_m - y(x_m)\| \leq C \left(\max_{0 \leq j < k} \|y_j - y(x_j)\| + h \max_{0 \leq j < k} \|f(x_j, y_j) - y'(x_j)\| \right) + Mh^p,$$

where the constants C and M are as in Theorem 8.2. \square

In the next two subsections we shall study the existence and construction of multipliers, and try to better understand the condition (8.22).

Construction of Multipliers. Obviously $\mu(\zeta) = 1$ is a multiplier iff the method itself is A -stable. Moreover, the limit $|\zeta| \rightarrow \infty$ in (8.20) shows that $\mu(0)$ must have the same sign as α_k/β_k (which we always assume to be positive). Therefore, the simplest (and most important) nontrivial multiplier has the form

$$\mu(\zeta) = 1 - \eta\zeta. \quad (8.23)$$

Suppose now that the method (ϱ, σ) is stable at infinity. The maximum principle for harmonic functions then implies that (8.23) is a multiplier for (ϱ, σ) iff $|\eta| \leq 1$ and

$$\operatorname{Re} \left((1 - \eta e^{it}) \frac{\varrho(e^{it})}{\sigma(e^{it})} \right) \geq 0 \quad \text{for all } t \in \mathbb{R}.$$

This condition motivates the study of

$$\gamma(t) = \left(\operatorname{Re} \left(\frac{\varrho(e^{it})}{\sigma(e^{it})} \right), -\operatorname{Re} \left(e^{it} \frac{\varrho(e^{it})}{\sigma(e^{it})} \right) \right), \quad (8.24)$$

which is called the *modified root-locus* curve by Nevanlinna & Odeh (1981). We then have:

Criterion 8.6. Consider a method which is stable at infinity. The function (8.23) is a *multiplier* for (ϱ, σ) iff $|\eta| \leq 1$ and the modified root-locus curve lies to the right of the straight line through the origin with slope $-1/\eta$.

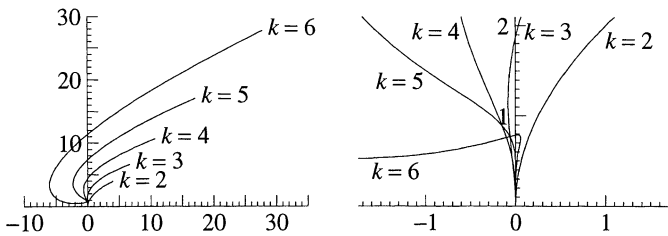


Fig. 8.1. Modified root-locus curve for BDF schemes

Fig. 8.1 shows the modified root-locus curves for the BDF schemes for $2 \leq k \leq 6$. The optimal values for η are given in Table 8.1.

Table 8.1. Multiplier for BDF schemes

k	η	$\arccos \eta$	$A(\alpha)$ -stable
2	0	$\pi/2$	$\pi/2$
3	0.0836	85.20°	86.03°
4	0.2878	73.27°	73.35°
5	0.8160	35.32°	51.84°
6	5.0130	—	17.84°

Proposition 8.7. If $\mu(\zeta)$ is a multiplier for (ϱ, σ) and we have

$$|\arg \mu(\zeta)| \leq \frac{\pi}{2} - \alpha \quad \text{for } |\zeta| \leq 1 \quad (8.25)$$

then the method is $A(\alpha)$ -stable.

Proof. Condition (8.20) together with (8.25) implies that

$$\left| \arg \left(\frac{\varrho(\zeta)}{\sigma(\zeta)} \right) - \pi \right| \geq \alpha \quad \text{for } |\zeta| \geq 1.$$

But this condition implies $A(\alpha)$ -stability. \square

A simple calculation shows that the multiplier (8.23) satisfies (8.25) with $\alpha = \arccos \eta$. For the BDF schemes we have included these values in Table 8.1 together with the α -values for linear stability.

Multipliers and Nonlinearities

We still have to investigate the problem under what conditions on the multiplier $\mu(\zeta)$ and on the function $f(x, y)$ one has (8.22) for all sequences $\{u_j\}$ and $\{v_j\}$. To get an idea of the nature of (8.22) we first look, following Nevanlinna & Odeh (1981), at the linear problem $y' = Ay$.

Proposition 8.8. *If the multiplier $\mu(\zeta)$ satisfies (8.25) and if the range of the matrix A lies in the sector $|\arg \langle Au, u \rangle - \pi| \leq \alpha$ for all $u \in \mathbb{C}^n$, then we have*

$$\sum_{m=0}^N \sum_{j=0}^m \mu_{m-j} \operatorname{Re} \langle Au_m, u_j \rangle \leq 0 \quad (8.26)$$

for all $N \geq 0$ and all sequences $\{u_j\}$.

Proof. A direct computation shows that the expression in (8.26) equals

$$\operatorname{Re} \left(\frac{1}{2\pi} \int_0^{2\pi} \mu(e^{it}) \langle A\hat{u}_N(t), \hat{u}_N(t) \rangle dt \right) \quad (8.27)$$

where

$$\hat{u}_N(t) = \sum_{j=0}^N e^{-ijt} u_j$$

denotes the Fourier transform of (u_0, u_1, \dots, u_N) . The assumptions on $\mu(\zeta)$ and on A imply that the integrand in (8.27) has non-positive real part. This proves (8.26). \square

Problems which satisfy (8.22) for some multiplier $\mu(\zeta)$ must also satisfy the one-sided Lipschitz condition (8.1) with $\nu = 0$ (this is seen by putting $N = 0$ in (8.22)). A class of nonlinear problems, for which (8.22) holds, is given by the following perturbation result.

Proposition 8.9. Let $f(x, y) = -Ay + Ag(x, y)$ where A is a symmetric and positive semi-definite matrix. With $\|u\|_A^2 = u^T Au$ suppose that

$$\|g(x, y) - g(x, z)\|_A \leq L\|y - z\|_A. \quad (8.28)$$

Then Condition (8.22) holds if

$$L \cdot \max_{|\zeta|=1} |\mu(\zeta)| \leq \min_{|\zeta|=1} \operatorname{Re} \mu(\zeta). \quad (8.29)$$

Remark. For the multiplier (8.23) Condition (8.29) is equivalent to $L \cdot (1 + \eta) \leq (1 - \eta)$.

Proof. As in the proof of Proposition 8.8 we get for $w_j = u_j - v_j$

$$\begin{aligned} - \sum_{m=0}^N \sum_{j=0}^m \mu_{m-j} \operatorname{Re} \langle Aw_m, w_j \rangle &= -\operatorname{Re} \left(\frac{1}{2\pi} \int_0^{2\pi} \mu(e^{it}) \langle A\widehat{w}_N(t), \widehat{w}_N(t) \rangle dt \right) \\ &\leq -m_0 \frac{1}{2\pi} \int_0^{2\pi} \langle A\widehat{w}_N(t), \widehat{w}_N(t) \rangle dt = -m_0 \sum_{j=0}^N \langle Aw_j, w_j \rangle \end{aligned} \quad (8.30)$$

where $m_0 = \min \operatorname{Re} \mu(e^{it})$. On the other hand, the inequality of Cauchy-Schwarz gives

$$\begin{aligned} \sum_{m=0}^N \operatorname{Re} \left\langle A(g(x_m, u_m) - g(x_m, v_m)), \sum_{j=0}^m \mu_{m-j}(u_j - v_j) \right\rangle \\ \leq \left(\sum_{m=0}^N \|g(x_m, u_m) - g(x_m, v_m)\|_A^2 \right)^{1/2} \cdot \left(\sum_{m=0}^N \left\| \sum_{j=0}^m \mu_{m-j}(u_j - v_j) \right\|_A^2 \right)^{1/2}. \end{aligned} \quad (8.31)$$

The last term in (8.31) can be estimated as (for the moment put $w_j = 0$ for $j > N$)

$$\begin{aligned} \sum_{m=0}^N \left\| \sum_{j=0}^m \mu_{m-j} w_j \right\|_A^2 &\leq \frac{1}{2\pi} \int_0^{2\pi} \left\| \sum_{m \geq 0} e^{-imt} \sum_{j=0}^m \mu_{m-j} w_j \right\|_A^2 dt \\ &= \frac{1}{2\pi} \int_0^{2\pi} |\mu(e^{-it})|^2 \cdot \left\| \sum_{j \geq 0} e^{-ijt} w_j \right\|_A^2 dt \leq M^2 \sum_{j=0}^N \|w_j\|_A^2 \end{aligned}$$

where $M = \max |\mu(e^{-it})|$. These estimates together with (8.28) show that the expression in (8.22) is majorized by

$$(L \cdot M - m_0) \sum_{j=0}^N \|u_j - v_j\|_A^2.$$

This is non-positive if (8.29) holds. \square

Discrete Variation of Constants and Perturbations

We now turn our attention to the perturbation approach of Lubich (1991), which extends the ideas of Sect. V.7 (discrete variation of constants) to nonlinear problems. For this we consider nonlinear differential equations written in the form

$$y' = Ay + g(t, y). \quad (8.32)$$

Inserting this equation into Formulas (8.2), (8.3), and (8.4) we get

$$\sum_{i=0}^k (\alpha_i I - hA\beta_i) \Delta y_{m+i} = h\Delta g_{m+k} + d_{m+k}, \quad (8.33)$$

where

$$\Delta g_{m+k} = \sum_{i=0}^k \beta_i \left(g(x_{m+i}, \hat{y}_{m+i}) - g(x_{m+i}, y_{m+i}) \right) \quad (8.34)$$

for $m \geq 0$. We further put $\Delta g_j = 0$ for $j < k$. Recall that d_j (for $j \geq k$) are usually the local truncation errors and d_0, \dots, d_{k-1} are defined by (8.33) with $m \in \{-1, \dots, -k\}$. The differences Δy_j are then the global errors of the method. If we introduce the formal power series

$$\Delta y(\zeta) = \sum_{j \geq 0} \Delta y_j \zeta^j, \quad \Delta g(\zeta) = \sum_{j \geq 0} \Delta g_j \zeta^j, \quad d(\zeta) = \sum_{j \geq 0} d_j \zeta^j$$

then the recursion (8.33) can be written as

$$\Delta y(\zeta) = r(\zeta, hA) (h\Delta g(\zeta) + d(\zeta)). \quad (8.35)$$

The resolvent $r(\zeta, hA)$ was introduced in (7.44) and (7.50). The coefficient of ζ^m in (8.35) then yields

$$\Delta y_m = h \sum_{j=0}^m r_{m-j}(hA) \Delta g_j + \sum_{j=0}^m r_{m-j}(hA) d_j. \quad (8.36)$$

The second sum on the right-hand side of (8.36) can be estimated as in Sect. V.7. In order to estimate the first term we have to combine estimates for $r_j(hA)$ with a Lipschitz condition for $g(x, y)$. This will lead to a Gronwall-type inequality, whose resolution gives the desired estimates for Δy_m . Let us illustrate this procedure in a simple situation.

Theorem 8.10. *Let the multistep method and the matrix A satisfy the assumptions of Theorem 7.10. If the nonlinearity $g(x, y)$ satisfies*

$$\|g(x, y) - g(x, z)\| \leq L\|y - z\| \quad (8.37)$$

then there exist constants C , h_0 and γ as in Theorem 7.10, and Λ (h_0 and Λ

depend on L) such that

$$\begin{aligned} & \|y(x_m) - y_m\| \\ & \leq C e^{\Lambda x_m} \left(\max_{0 \leq j < k} \|y(x_j) - y_j\| + h^p \int_{x_0}^{x_m} e^{\gamma(x_m - \xi)} \|y^{(p+1)}(\xi)\| d\xi \right). \end{aligned}$$

Proof. It follows from the proof of Theorem 7.10 and from (8.36) that

$$\|\Delta y_m\| \leq h L C_1 \sum_{j=0}^m e^{\gamma(m-j)h} \|\Delta y_j\| + C_2 \sum_{j=0}^m e^{\gamma(m-j)h} \varepsilon_j, \quad (8.38)$$

where (with $0 \leq \kappa < 1$)

$$\varepsilon_m = C_0 \left(\kappa^m \max_{0 \leq j < k} \|\Delta y_j\| + h^p \int_{x_{m-k}}^{x_m} \|y^{(p+1)}(\xi)\| d\xi \right).$$

Application of Exercise 1 to the sequence $\{e^{-\gamma m h} \|\Delta y_m\|\}$ yields the statement of the theorem. \square

Lubich (1991) has shown how the above estimates can be improved to obtain convergence results for singularly perturbed problems (see Sect. VI.2) and for discretized nonlinear parabolic equations, as we shall see in the sequel.

Convergence for Nonlinear Parabolic Problems

We consider the initial value problem

$$y' + Ay = g(t, y), \quad y(0) = y_0 \quad (8.39)$$

obtained by space discretization of a parabolic differential equation. The matrix A is assumed to satisfy for some $\alpha' \in (0, \pi/2)$

$$\|(sI + A)^{-1}\| \leq \frac{M}{1 + |s|} \quad \text{for} \quad |\arg s| \leq \pi - \alpha' \quad (8.40)$$

(compare (7.47)). In order to motivate our assumptions on $g(t, y)$ we begin with two examples.

Burgers' Equation. For this problem (Burgers 1948)

$$u_t + uu_x = \mu u_{xx} \quad \text{or} \quad u_t + \left(\frac{u^2}{2}\right)_x = \mu u_{xx}, \quad \mu > 0,$$

we consider the discretization

$$\dot{u}_i = -\frac{u_{i+1}^2 - u_{i-1}^2}{4\Delta x} + \mu \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta x)^2}.$$

It is of the form (8.39) with

$$A = \frac{\mu}{\Delta x^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix}, \quad g(t, y) = \frac{1}{4\Delta x} \begin{pmatrix} y_2^2 - y_0^2 \\ y_3^2 - y_1^2 \\ \vdots \\ y_{n+1}^2 - y_{n-1}^2 \end{pmatrix}, \quad (8.41)$$

where $\mu > 0$ is a given constant, $\Delta x = 1/(n+1)$ and, due to homogeneous boundary conditions, $y_0 = y_{n+1} = 0$. In this situation we work with the scaled norm (on \mathbb{R}^n)

$$\|u\| = \sqrt{\Delta x \sum_{i=1}^n |u_i|^2}, \quad (8.42)$$

which tends to that of $L^2(0, 1)$ for $n \rightarrow \infty$. As the eigenvalues of the symmetric matrix A are real and positive, condition (8.40) is verified for every $\alpha' > 0$, uniformly in Δx .

The presence of the denominator Δx in $g(t, y)$ of (8.41) does not allow a Lipschitz condition (8.37) uniformly in $\Delta x > 0$ (not even in a neighbourhood of the exact solution). However, using the energy norm $\|A^{1/2}u\|$, which already contains the factor $1/\Delta x$, we show that

$$\|g(t, y) - g(t, z)\| \leq \mu^{-1} r \cdot \|A^{1/2}(y - z)\| \quad \text{for } \|A^{1/2}y\| + \|A^{1/2}z\| \leq r. \quad (8.43)$$

For the proof of this relation we consider the bilinear map $b : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n$, whose i th component is defined by

$$b_i(u, v) = (4\Delta x)^{-1} (u_{i+1} + u_{i-1})(v_{i+1} - v_{i-1})$$

(again we put $u_0 = v_0 = u_{n+1} = v_{n+1} = 0$). Then

$$g(t, y) - g(t, z) = b(y, y) - b(z, z) = b(y, y - z) + b(y - z, z), \quad (8.44)$$

and we need an estimate for $\|b(u, v)\|$. Using

$$|(u_{i+1} + u_{i-1})(v_{i+1} - v_{i-1})| \leq 2 \cdot |v_{i+1} - v_{i-1}| \cdot \max_j |u_j|,$$

and the estimates of Exercise 3 we obtain

$$\|b(u, v)\| \leq \|u\|_\infty \cdot \|Dv\| \leq \mu^{-1} \cdot \|A^{1/2}u\| \cdot \|A^{1/2}v\|. \quad (8.45)$$

where

$$D = \frac{1}{2\Delta x} \begin{pmatrix} 0 & 1 & & & \\ -1 & 0 & 1 & & \\ & -1 & 0 & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & -1 & 0 \end{pmatrix} \quad (8.46)$$

represents the first central difference operator. The estimate (8.45) applied to (8.44) proves (8.43).

Incompressible Navier-Stokes Equation. The motion of a viscous incompressible fluid in a domain $\Omega \subset \mathbb{R}^d$ is governed by the equations of Navier (1823) and Stokes (1845)

$$\frac{\partial \mathbf{u}}{\partial t} + \sum_{i=1}^d u_i \frac{\partial \mathbf{u}}{\partial x_i} = \Delta \mathbf{u} - \mathbf{grad} p, \quad \operatorname{div} \mathbf{u} = 0 \quad (8.47)$$

where $\mathbf{u} = (u_1, \dots, u_d)^T$. We denote by P the orthogonal projection from $L^2(\Omega)^d$ onto X , where X is the subspace of functions with $\operatorname{div} \mathbf{u} = 0$ (more precisely: the closure of the set of smooth functions with vanishing divergence and support contained in Ω). If we apply P to Eq. (8.47), $\mathbf{grad} p$ is eliminated and we obtain

$$\frac{\partial \mathbf{u}}{\partial t} - P \Delta \mathbf{u} = -P \left(\sum_{i=1}^d u_i \frac{\partial \mathbf{u}}{\partial x_i} \right). \quad (8.48)$$

These equations are now precisely of the form (8.39), where $A = -P \Delta$ (or some discretization of it) and $g(t, y)$ is the right-hand side of (8.48). Lipschitz estimations for this nonlinear term have been obtained by Sobolevskii (1959) and Fujita & Kato (1964). They are of the form

$$\|g(t, u) - g(t, v)\|_{\beta-\gamma} \leq \ell(r) \cdot \|u - v\|_{\beta} \quad \text{for } \|u\|_{\beta} + \|v\|_{\beta} \leq r \quad (8.49)$$

where $\|\cdot\|_{\beta}$ denotes the norm

$$\|u\|_{\beta} = \|A^{\beta} u\|. \quad (8.50)$$

In particular, for $d = 3$, condition (8.49) is true for $\beta = 1/2$, $\gamma \geq 3/4$ as well as for $\beta = \gamma > 3/4$ (Fujita & Kato 1964, pp. 272-273).

Motivated by these examples we consider the initial value problem (8.39) on \mathbb{R}^n , where A is supposed to satisfy (8.40) for some $\alpha' \in (0, \pi/2)$ and the nonlinearity $g(t, y)$ is assumed to satisfy the Lipschitz condition (8.49).

Application of a linear multistep method to (8.39) yields

$$\sum_{i=0}^k \alpha_i y_{m+i} + hA \sum_{i=0}^k \beta_i y_{m+i} = h \sum_{i=0}^k \beta_i g(t_{m+i}, y_{m+i}). \quad (8.51)$$

Instead of comparing the numerical solution $\{y_m\}$ with the analytic solution $y(t)$ of (8.39), it is more interesting to compare it with the exact solution of the original partial differential equation. We therefore denote by $\eta(t)$ a projection of the solution of the PDE into the finite-dimensional space under consideration. In this way we obtain

$$\eta' + A\eta = g(t, \eta) + s(t)$$

where $s(t)$ is the spatial discretization error.

Theorem 8.11 (Lubich 1991). *Consider the problem (8.39) with A and $g(t, y)$ satisfying (8.40) and (8.49) with $\gamma < 1$, respectively. Assume that the multistep method is of order p , $A(\alpha)$ -stable for some $\alpha > \alpha'$, and strictly stable at infinity. Then, the full discretization error is bounded by*

$$\begin{aligned} \|y_m - \eta(t_m)\|_\beta &\leq C \cdot \left(\max_{0 \leq j < k} \|y_j - \eta(t_j)\|_\beta + h^p \int_0^{t_m} \|\eta^{(p+1)}(t)\|_\beta dt \right. \\ &\quad \left. + \|A^{-1}s(0)\|_\beta + \int_0^{t_m} \|A^{-1}s'(t)\|_\beta dt \right). \end{aligned} \quad (8.52)$$

The estimate holds for $t_m = mh \leq T$ provided that $h \leq h_0$ and the expression in brackets on the right-hand side is bounded by ε , where h_0 and ε are sufficiently small. The constants C , h_0 and ε depend on $\max_{0 \leq t \leq T} \|\eta(t)\|_\beta$ and M of (8.40), but are otherwise independent of A and the dimension of the system, and independent of m and h .

Proof. a) The projected solution $\eta(t)$ of the PDE, inserted into (8.51), gives

$$\sum_{i=0}^k \alpha_i \eta(t_{m+i}) = h \sum_{i=0}^k \beta_i \left(-A\eta(t_{m+i}) + g(t_{m+i}, \eta(t_{m+i})) + s(t_{m+i}) \right) + d_{m+k}$$

where

$$\|d_{m+k}\|_\beta \leq C_0 h^p \int_{t_m}^{t_{m+k}} \|\eta^{(p+1)}(t)\|_\beta dt, \quad m \geq 0. \quad (8.53)$$

The same analysis which was necessary for (8.36) now gives for the error $\Delta y_m = \eta(t_m) - y_m$ the relation

$$\Delta y_m = h \sum_{j=0}^m r_{m-j}(-hA) \Delta g_j + h \sum_{j=0}^m r_{m-j}(-hA) \Delta s_j + h \sum_{j=0}^m r_{m-j}(-hA) d_j. \quad (8.54)$$

As in (8.34) the quantities Δg_j and Δs_j are defined by

$$\begin{aligned} \Delta g_{m+k} &= \sum_{i=0}^k \beta_i \left(g(t_{m+i}, \eta(t_{m+i})) - g(t_{m+i}, y_{m+i}) \right) \\ \Delta s_{m+k} &= \sum_{i=0}^k \beta_i s(t_{m+i}) \end{aligned}$$

for $m \geq 0$, and $\Delta g_j = 0$, $\Delta s_j = 0$ for $j < k$. The values d_0, \dots, d_{k-1} are defined as usual (see their definition before (8.4')). The following three parts of the proof treat the three terms in the right-hand side of (8.54) separately.

b) The Lipschitz condition (8.49) can be written as

$$\|A^{-\gamma}(g(t, y) - g(t, z))\|_\beta \leq \ell(r) \cdot \|y - z\|_\beta \quad \text{for} \quad \|y\|_\beta + \|z\|_\beta \leq r.$$

We put $\varrho = \max_{0 \leq t \leq T} \|\eta(t)\|_\beta$ and assume that for $hm \leq T$ the numerical solution y_m exists and is bounded by $\|y_m\| \leq \varrho + 1$ (this will be verified recursively in part

(f) of the proof) so that

$$\|A^{-\gamma} \Delta g_{m+k}\|_{\beta} \leq \ell(2\varrho + 1) \cdot \sum_{i=0}^k |\beta_i| \cdot \|\Delta y_{m+i}\|_{\beta}. \quad (8.55)$$

Consequently we have to find an estimate for $\|r_{m-j}(-hA)A^{\gamma}\|_{\beta}$ (for the matrix norm corresponding to the vector norm $\|\cdot\|_{\beta}$; see Sect. I.9). We note that $\|r_{m-j}(-hA)A^{\gamma}\|_{\beta} = \|A^{\gamma}r_{m-j}(-hA)\|$ and recall that $A^{\gamma}r_j(-hA)$ is the coefficient of ζ^j in the series for

$$A^{\gamma}r(\zeta, -hA) = A^{\gamma}(\delta(\zeta)I + hA)^{-1} \frac{\zeta^{-k}}{\sigma(\zeta^{-1})}.$$

In order to apply Lemma 7.11 we have to estimate $\Phi(s) = A^{\gamma}(sI + A)^{-1}$. If A can be transformed to diagonal form with an orthogonal matrix (as it is the case for (8.41)), we have for $|\arg s| \leq \pi - \alpha'$ ($0 < \alpha' < \alpha$)

$$\|A^{\gamma}(sI + A)^{-1}\| \leq \sup_{a \geq 0} \frac{a^{\gamma}}{|s+a|} \leq M_1 \cdot |s|^{\gamma-1}.$$

For the general case we refer the reader to Henry (1981, pp. 26-28). Application of Lemma 7.11 (see also Remark 7.12) yields

$$\|r_j(-hA)A^{\gamma}\|_{\beta} \leq C_1((j+1)h)^{-\gamma} \quad \text{for } j \geq 0.$$

Together with the Lipschitz condition (8.55) this gives with $L = C_1 \cdot \ell(2\varrho + 1)$

$$h \left\| \sum_{j=0}^m r_{m-j}(-hA) \Delta g_j \right\|_{\beta} \leq h^{1-\gamma} L \sum_{j=0}^m (m-j+1)^{-\gamma} \|\Delta y_j\|_{\beta}. \quad (8.56)$$

c) The second term in (8.54) is the coefficient of ζ^m in

$$hr(\zeta, -hA)\Delta s(\zeta) = \widetilde{r}(\zeta)\widetilde{\Delta s}(\zeta)$$

where we have introduced

$$\begin{aligned} \widetilde{r}(\zeta) &= (\delta(\zeta)I + hA)^{-1} hA \delta(\zeta)^{-1} \frac{\zeta^{-k}}{\sigma(\zeta^{-1})} = \sum_{j \geq 0} \widetilde{r}_j \zeta^j \\ \widetilde{\Delta s}(\zeta) &= \delta(\zeta) A^{-1} \Delta s(\zeta) = \sum_{j \geq 0} \widetilde{\Delta s}_j \zeta^j. \end{aligned}$$

In order to estimate $\|\widetilde{r}_j\|_{\beta}$ (matrix norm) we note that $\|\widetilde{r}_j\|_{\beta} = \|\widetilde{r}_j\|$. In view of an application of Lemma 7.11 we have to consider $\Phi(s) = (sI + A)^{-1} A s^{-1} = s^{-1}I - (sI + A)^{-1}$ which, because of (8.40), is bounded by $(M+1)/|s|$. Lemma 7.11 thus yields $\|\widetilde{r}_j\|_{\beta} \leq C_2$. Further we have

$$\widetilde{\Delta s}(\zeta) = \frac{\delta(\zeta)}{1-\zeta} \cdot \left(A^{-1} \Delta s_k \zeta^k + \sum_{j \geq k+1} A^{-1} (\Delta s_j - \Delta s_{j-1}) \zeta^j \right)$$

where the coefficients of $\delta(\zeta)/(1-\zeta)$ are absolutely summable, because the zeros of $\sigma(\zeta)$ lie all inside $|\zeta| < 1$. Combining all these estimates we get

$$\begin{aligned} h \left\| \sum_{j=0}^m r_{m-j}(-hA) \Delta s_j \right\|_{\beta} &= \left\| \sum_{j=0}^m \tilde{r}_{m-j} \widetilde{\Delta s_j} \right\|_{\beta} \\ &\leq C_3 \left(\|A^{-1} \Delta s_k\|_{\beta} + \sum_{j=k+1}^m \|A^{-1} (\Delta s_j - \Delta s_{j-1})\|_{\beta} \right) \\ &\leq C_4 \left(\|A^{-1} s(0)\|_{\beta} + \int_0^{t_m} \|A^{-1} s'(t)\|_{\beta} dt \right). \end{aligned} \quad (8.57)$$

d) The last term in (8.54) can be estimated in the same way as the corresponding term in the proof of Theorem 7.10. We just have to take the norm (8.50) and get

$$h \left\| \sum_{j=0}^m r_{m-j}(-hA) d_j \right\|_{\beta} \leq C_5 \left(\max_{0 \leq j < k} \|y_j - \eta(t_j)\|_{\beta} + h^p \int_0^{t_m} \|\eta^{(p+1)}(t)\|_{\beta} dt \right). \quad (8.58)$$

e) Inserting (8.56), (8.57), and (8.58) into (8.54) gives

$$\|\Delta y_m\|_{\beta} \leq h^{1-\gamma} L \sum_{j=0}^m (m-j+1)^{-\gamma} \|\Delta y_j\|_{\beta} + C_6 \varepsilon_m \quad (8.59)$$

where $C_6 = \max(C_4, C_5)$ and ε_m denotes the expression in brackets on the right-hand side of (8.52). For $h \leq h_0$ and $h_0^{1-\gamma} L < 1$ this Gronwall-type inequality can be solved (Exercise 2) and gives $\|\Delta y_m\|_{\beta} \leq C_7 \varepsilon_m$, the desired result.

f) We now justify recursively our assumption $\|y_m\|_{\beta} \leq \varrho + 1$ used in (b). Suppose that $\|y_j\|_{\beta} \leq \varrho + 1$ for $j = 0, 1, \dots, m-1$, then it follows from $h^{1-\gamma} L < 1$ and the contraction mapping theorem that a unique solution y_m of (8.54) exists. This solution verifies $\|y_m\|_{\beta} \leq \|\eta(t_m)\|_{\beta} + \|\Delta y_m\|_{\beta} \leq \varrho + 1$ if ε is small enough, more precisely, if $C_7 \varepsilon < 1$. \square

Remark. A different approach to convergence results of multistep methods for nonlinear parabolic equations is given by Le Roux (1980). A corresponding theorem for Runge-Kutta methods is proved in Lubich & Ostermann (1993).

Exercises

1. Let $L \geq 0$ and consider two sequences $\{u_j\}$ and $\{\varepsilon_j\}$ of nonnegative numbers which satisfy

$$u_m \leq hL \sum_{j=0}^m u_j + \sum_{j=0}^m \varepsilon_j \quad \text{for } m \geq 0.$$

Prove that for $hL \leq 1 - C^{-1}$

$$u_m \leq C e^{LCmh} \sum_{j=0}^m \varepsilon_j.$$

Hint. Show by induction that $v_m \leq h\Lambda \sum_{j=0}^{m-1} v_j + M$ implies $v_m \leq M(1 + h\Lambda)^m \leq M e^{\Lambda mh}$.

2. Consider the inequality (8.59) with $\gamma < 1$, $L \geq 0$, $\varepsilon_m \geq 0$ and $h > 0$. Under the assumptions $h \leq h_0$ and $h_0^{1-\gamma}L < 1$ prove that there exists a constant C such that $\|\Delta y_m\|_\beta \leq C\varepsilon_m$ for $mh \leq T$.

Hint. Move the term $h^{1-\gamma}L\|\Delta y_m\|_\beta$ to the left and divide the inequality by $(1 - h^{1-\gamma}L)$. This yields

$$\|\Delta y_m\|_\beta \leq h^{1-\gamma} \hat{L} \sum_{j=0}^{m-1} (m-j)^{-\gamma} \|\Delta y_j\|_\beta + \hat{\varepsilon} \quad \text{for } m \geq 0.$$

Show that $\|\Delta y_m\|_\beta \leq \hat{\varepsilon} u(mh)$, where $u(x)$ is the solution of

$$u(x) = 1 + \hat{L} \int_0^x (x-t)^{-\gamma} u(t) dt. \quad (8.60)$$

Estimate the solution of (8.60) (see Henry 1981, pp. 188-190).

3. Let A and D be the matrices of (8.41) (suppose $\mu = 1$) and (8.46). Prove that for all $u \in \mathbb{R}^n$

$$\text{a) } \|u\|_\infty \leq \|A^{1/2}u\|, \quad \text{b) } \|Du\| \leq \|A^{1/2}u\|,$$

where $\|u\|_\infty = \max_i |u_i|$ and $\|\cdot\|$ is the norm of (8.42).

Hint. a) Let $u_0 = 0$ and apply the inequality of Cauchy-Bunyakovski-Schwarz to $u_i = \sum_{j=1}^i (u_j - u_{j-1})$. This gives

$$\|u\|_\infty^2 \leq \sum_{j=1}^n \left(\frac{u_j - u_{j-1}}{\Delta x} \right)^2 = u^T A u.$$

b) The inequality $u^T A u \geq \|Du\|^2$ is a consequence of the algebraic identity ($u_0 = u_{n+1} = 0$)

$$\begin{aligned} 4 \sum_{i=1}^n (2u_i^2 - u_i u_{i+1} - u_i u_{i-1}) - \sum_{i=1}^n (u_{i+1} - u_{i-1})^2 \\ = \sum_{i=1}^n (u_{i+1} - 2u_i + u_{i-1})^2 + 2u_1^2 + 2u_n^2. \end{aligned}$$

V.9 Algebraic Stability of General Linear Methods

General linear methods were originally introduced as a means of unifying and generalizing existing theories for traditional methods. (J.C. Butcher 1987)

In Sections IV.12 and V.6 we have studied the nonlinear stability of Runge-Kutta methods (B -stability) and of one-leg methods (G -stability). It is natural to ask whether these theories can be combined within the class of general linear methods. This work was initiated by Burrage & Butcher (1980).

We consider the differential equation $y' = f(x, y)$ where y and f are complex-valued vectors and we assume the one-sided Lipschitz condition

$$\operatorname{Re} \langle f(x, y) - f(x, z), y - z \rangle \leq \nu \|y - z\|^2. \quad (9.1)$$

General linear methods are defined by (see Example 8.5 of Sect. III.8)

$$u_i^{(n+1)} = \sum_{j=1}^k a_{ij} u_j^{(n)} + h \sum_{j=1}^s b_{ij} f(x_n + c_j h, v_j^{(n)}), \quad i = 1, \dots, k \quad (9.2a)$$

$$v_i^{(n)} = \sum_{j=1}^k \tilde{a}_{ij} u_j^{(n)} + h \sum_{j=1}^s \tilde{b}_{ij} f(x_n + c_j h, v_j^{(n)}), \quad i = 1, \dots, s. \quad (9.2b)$$

Here, $u_n = (u_1^{(n)}, \dots, u_k^{(n)})^T$ contains the necessary information from the previous step. The internal stages $(v_1^{(n)}, \dots, v_s^{(n)})$, defined by (9.2b), serve for the computation of u_{n+1} in (9.2a).

G -Stability

As in Sect. V.6, we consider inner product norms

$$\|u_n\|_G^2 = \sum_{i=1}^k \sum_{j=1}^k g_{ij} \langle u_i^{(n)}, u_j^{(n)} \rangle, \quad (9.3)$$

where $G = (g_{ij})$ is a real, symmetric and positive definite matrix.

Definition 9.1. The general linear method (9.2) is called G -stable, if there exists a real, symmetric and positive definite matrix G , such that for two numerical solutions $\{u_n\}$ and $\{\hat{u}_n\}$,

$$\|u_{n+1} - \hat{u}_{n+1}\|_G \leq \|u_n - \hat{u}_n\|_G \quad (9.4)$$

for all step sizes $h > 0$ and for all differential equations satisfying (9.1) with $\nu = 0$.

For Runge-Kutta methods (where $k = 1$ and apart from a scaling factor $G = (1)$) this definition reduces to B -stability as introduced in Definition IV.12.2. For one-leg methods (where $s = 1$ and $u_n = (y_{n+k-1}, \dots, y_n)^T$) it is equivalent to Definition 6.3.

Many methods can be written in different ways as general linear methods and the above definition of G -stability may depend on the particular formulation. For example, the trapezoidal rule

$$y_{n+1} = y_n + \frac{h}{2}(f(x_n, y_n) + f(x_{n+1}, y_{n+1}))$$

can be considered as a Runge-Kutta method (with $u_n = y_n$). In this case it is not G -stable (because it is not B -stable, see Theorem IV.12.12). However, if we let $u_n = (y_n, hy'_n)$ where $y'_n = f(x_n, y_n)$, then the trapezoidal rule satisfies (9.4) with

$$G = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/4 \end{pmatrix}. \quad (9.5)$$

This follows from the fact that whenever $\{y_n\}$ is the solution obtained by the trapezoidal rule, then $z_n = y_n + \frac{h}{2}y'_n$ is a solution of the implicit midpoint rule, which is known to be B -stable (see Example IV.12.3 or Theorem IV.12.9). Therefore

$$\|y_{n+1} + \frac{h}{2}y'_{n+1}\| \leq \|y_n + \frac{h}{2}y'_n\|$$

which proves the statement. The matrix G in (9.5) is singular and thus not strictly positive definite. Burrage & Butcher (1980), however, admit non-zero non-negative definite matrices G in their definition of G -stability (which they call *monotonicity*). Therefore the trapezoidal rule is G -stable in their definition.

Algebraic Stability

In addition to (9.2) we consider a second numerical solution (marked with hats) produced by the same method using different starting values. We denote the differences by

$$\begin{aligned} \Delta u_i^{(n)} &= u_i^{(n)} - \hat{u}_i^{(n)}, & \Delta u_n &= u_n - \hat{u}_n \\ \Delta v_i^{(n)} &= v_i^{(n)} - \hat{v}_i^{(n)}, & \Delta f_i^{(n)} &= h(f(x_n + c_i h, v_i^{(n)}) - f(x_n + c_i h, \hat{v}_i^{(n)})). \end{aligned}$$

The following lemma states an identity which will be essential in the study of G -stability.

Lemma 9.2 (Burrage & Butcher 1980). *Let G be a real, symmetric matrix and $D = \text{diag}(d_1, \dots, d_s)$ be a real diagonal matrix. The difference of two solutions of (9.2) then satisfies*

$$\|\Delta u_{n+1}\|_G^2 - \|\Delta u_n\|_G^2 = 2 \sum_{i=1}^s d_i \text{Re} \langle \Delta f_i^{(n)}, \Delta v_i^{(n)} \rangle - \sum_{i,j=1}^{s+k} m_{ij} \langle w_i, w_j \rangle$$

where $(w_1, \dots, w_{s+k}) = (\Delta u_1^{(n)}, \dots, \Delta u_k^{(n)}, \Delta f_1^{(n)}, \dots, \Delta f_s^{(n)})$ and the matrix $M = (m_{ij})$ is given by

$$M = \begin{pmatrix} G - A^T G A & \tilde{A}^T D - A^T G B \\ D \tilde{A} - B^T G A & D \tilde{B} + \tilde{B}^T D - B^T G B \end{pmatrix}. \quad (9.6)$$

Proof. We consider the identity

$$\begin{aligned} & \|\Delta u_{n+1}\|_G^2 - \|\Delta u_n\|_G^2 - 2 \sum_{i=1}^s d_i \operatorname{Re} \langle \Delta f_i^{(n)}, \Delta v_i^{(n)} \rangle \\ &= \sum_{i,j=1}^k g_{ij} \langle \Delta u_i^{(n+1)}, \Delta u_j^{(n+1)} \rangle - \sum_{i,j=1}^k g_{ij} \langle \Delta u_i^{(n)}, \Delta u_j^{(n)} \rangle \\ & \quad - \sum_{i=1}^s d_i \langle \Delta f_i^{(n)}, \Delta v_i^{(n)} \rangle - \sum_{i=1}^s d_i \langle \Delta v_i^{(n)}, \Delta f_i^{(n)} \rangle \end{aligned}$$

and insert the formulas (9.2). This gives

$$\begin{aligned} \dots &= \sum_{i,j=1}^k g_{ij} \left\langle \sum_{\ell=1}^k a_{i\ell} \Delta u_\ell^{(n)} + h \sum_{\ell=1}^s b_{i\ell} \Delta f_\ell^{(n)}, \sum_{\ell=1}^k a_{j\ell} \Delta u_\ell^{(n)} + h \sum_{\ell=1}^s b_{j\ell} \Delta f_\ell^{(n)} \right\rangle \\ & \quad - \sum_{i,j=1}^k g_{ij} \langle \Delta u_i^{(n)}, \Delta u_j^{(n)} \rangle - \sum_{i=1}^s d_i \left\langle \Delta f_i^{(n)}, \sum_{\ell=1}^k \tilde{a}_{i\ell} \Delta u_\ell^{(n)} + h \sum_{\ell=1}^s \tilde{b}_{i\ell} \Delta f_\ell^{(n)} \right\rangle \\ & \quad - \sum_{i=1}^s d_i \left\langle \sum_{\ell=1}^k \tilde{a}_{i\ell} \Delta u_\ell^{(n)} + h \sum_{\ell=1}^s \tilde{b}_{i\ell} \Delta f_\ell^{(n)}, \Delta f_i^{(n)} \right\rangle. \end{aligned}$$

Multiplying out and collecting suitable terms proves the statement. \square

Definition 9.3. The general linear method (9.2) is called *algebraically stable*, if there exist a real, symmetric and positive definite matrix G and a real non-negative definite diagonal matrix D , such that the matrix M of (9.6) is non-negative definite.

An immediate consequence of our assumption (9.1) with $\nu = 0$ and of Lemma 9.2 is the following result.

Theorem 9.4. *Algebraic stability implies G -stability.* \square

For a given method it may be difficult to find matrices D and G such that M of (9.6) is non-negative definite. The following lemma shows some useful relations,

which hold if the method is assumed to be *preconsistent*, i.e., if there exists a vector $\xi_0 \in \mathbb{R}^k$ such that

$$A\xi_0 = \xi_0, \quad \tilde{A}\xi_0 = \mathbb{1} \quad (9.7)$$

(cf. Eq. (8.25) of Sect. III.8).

Lemma 9.5. *If a general linear method is preconsistent and algebraically stable, then the matrices D and G satisfy*

- i) $(d_1, \dots, d_s)^T = D\mathbb{1} = B^T G\xi_0$,
- ii) $(I - A^T)G\xi_0 = 0$, i.e., $G\xi_0$ is a left-eigenvector of A corresponding to the eigenvalue 1.

Proof. i) Let $\eta \in \mathbb{R}^s$ and $\varepsilon \in \mathbb{R}$ be arbitrary. The non-negativity of M , given by (9.6), implies

$$(\xi_0^T, \varepsilon\eta^T)M \begin{pmatrix} \xi_0 \\ \varepsilon\eta \end{pmatrix} \geq 0$$

so that

$$\xi_0^T (G - A^T G A) \xi_0 + 2\varepsilon\eta^T (D\tilde{A} - B^T G A) \xi_0 + \varepsilon^2 \eta^T (D\tilde{B} + \tilde{B}^T D - B^T G B) \eta \geq 0.$$

Since the ε -independent term vanishes (due to $A\xi_0 = \xi_0$), the coefficient of ε must be zero and since this holds for all η , the result follows.

ii) A similar argument applied to

$$(\xi_0 + \varepsilon\xi_1)^T (G - A^T G A) (\xi_0 + \varepsilon\xi_1) \geq 0 \quad \text{for all } \xi_1 \in \mathbb{R}^k, \varepsilon \in \mathbb{R}$$

implies the second statement. \square

AN-Stability and Equivalence Results

It is interesting to study in which situation algebraic stability is also necessary for G -stability. For this we consider the differential equation

$$y' = \lambda(x)y \quad \text{with} \quad \operatorname{Re} \lambda(x) \leq 0.$$

If we apply the general linear method (9.2) to this problem, we obtain

$$u_{n+1} = S(Z)u_n \quad (9.8)$$

where $Z = \operatorname{diag}(z_1, \dots, z_s)$, $z_j = h\lambda(x_n + c_j h)$ and

$$S(Z) = A + BZ(I - \tilde{B}Z)^{-1}\tilde{A}. \quad (9.9)$$

In the sequel we assume that the abscissae c_j are related to the other coefficients of the method by (see also Remark III.8.17)

$$(c_1, \dots, c_s)^T = c = \tilde{A}\xi_1 + \tilde{B}\mathbb{1}, \quad (9.10)$$

where $\xi_1 \in \mathbb{R}^k$ is the second coefficient vector of the exact value function

$$z(x, h) = y(x)\xi_0 + hy'(x)\xi_1 + \mathcal{O}(h^2).$$

This means that the internal stages approximate the exact solution as

$$v_j^{(n)} = y(x_n + c_j h) + \mathcal{O}(h^2).$$

Definition 9.6. A general linear method is called *AN-stable*, if there exists a real, symmetric and positive definite matrix G such that

$$\|S(Z)u\|_G \leq \|u\|_G \quad \text{for all } Z = \text{diag}(z_1, \dots, z_s) \text{ satisfying } \text{Re } z_j \leq 0 \\ (j = 1, \dots, s) \text{ and } z_j = z_k \text{ whenever } c_j = c_k.$$

Other possible definitions of *AN-stability* are given in Butcher (1987). For example, if the condition $\|S(Z)u\|_G \leq \|u\|_G$ is replaced by the powerboundedness of the matrix $S(Z)$, the method is called *weakly AN-stable*. This definition, however, does not allow the values $z_j = h\lambda(x_n + c_j h)$ to change at each step. Another modification is to consider arbitrary norms (instead of inner product norms only) in the definition of *AN-stability*. Butcher (1987) has shown that this does not lead to a larger class of *AN-stable* methods, but makes the analysis much more difficult.

We are now interested in the relations between the various stability definitions: the implications

$$\text{algebraically stable} \implies G\text{-stable} \implies AN\text{-stable} \implies A\text{-stable}$$

are either trivial or follow from Theorem 9.4. We also know that *A-stability* does not, in general, imply *AN-stability* (see e.g., Theorem IV.12.12). The following result shows that the other two implications are (nearly always) reversible.

Theorem 9.7 (Butcher 1987). *For preconsistent and non-confluent general linear methods (i.e., methods with distinct c_j) we have*

$$\text{algebraically stable} \iff G\text{-stable} \iff AN\text{-stable}.$$

Proof. It is sufficient to prove that *AN-stability* implies algebraic stability. For this we take the matrix G , whose existence is known by the definition of *AN-stability*, and show that the matrices D and M , given by Lemma 9.5i and (9.6), are non-negative definite.

In order to prove $d_j \geq 0$ we put $z_j = -\varepsilon$ ($\varepsilon > 0$) and $z_k = 0$ for $k \neq j$. We further let $\Delta u_n = \xi_0$ (the preconsistency vector of (9.7)) and $\Delta f_\ell^{(n)} = z_\ell \Delta v_\ell^{(n)}$, so that $\Delta u_{n+1} = S(Z)\xi_0$ and $\Delta v_\ell^{(n)} = 1 + \mathcal{O}(\varepsilon)$. Using

$$M \begin{pmatrix} \xi_0 \\ 0 \end{pmatrix} = 0, \tag{9.11}$$

which is a consequence of Lemma 9.5, the identity of Lemma 9.2 yields

$$\|S(Z)\xi_0\|_G^2 - \|\xi_0\|_G^2 = -2\varepsilon d_j + \mathcal{O}(\varepsilon^2).$$

Since the left-hand side of this equation is non-positive by AN -stability, we obtain $d_j \geq 0$.

We next put $z_\ell = i\varepsilon\eta_\ell$ where $\eta = (\eta_1, \dots, \eta_s)^T \in \mathbb{R}^s$ is arbitrary and ε is a small real parameter. We further put $\Delta u_n = \xi_0 + i\varepsilon\mu$ with $\mu \in \mathbb{R}^k$ and $\Delta f_\ell^{(n)} = z_\ell \Delta v_\ell^{(n)}$. This again implies $\Delta v_\ell^{(n)} = 1 + \mathcal{O}(\varepsilon)$. The identity of Lemma 9.2 together with (9.11) gives

$$\begin{aligned} \|S(Z)\xi_0\|_G^2 - \|\xi_0\|_G^2 &= -(\xi_0 - i\varepsilon\mu, i\varepsilon\eta + \mathcal{O}(\varepsilon^2))M \begin{pmatrix} \xi_0 + i\varepsilon\mu \\ i\varepsilon\eta + \mathcal{O}(\varepsilon^2) \end{pmatrix} = \\ &= -\varepsilon^2(\mu, \eta)^T M \begin{pmatrix} \mu \\ \eta \end{pmatrix} + \mathcal{O}(\varepsilon^3). \end{aligned}$$

Since this relation holds for all μ and η , the matrix M has to be non-negative definite. \square

Example 9.8. Let us investigate the G -stability of *multistep collocation methods* as introduced in Sect. V.3. We consider here the case $k = 2$ and $s = 2$, and fix one collocation point at $c_2 = 1$. The method is then given by

$$\begin{aligned} \begin{pmatrix} y_{n+1} \\ y_n \end{pmatrix} &= \overbrace{\begin{pmatrix} 1 - \varphi(1) & \varphi(1) \\ 1 & 0 \end{pmatrix}}^A \begin{pmatrix} y_n \\ y_{n-1} \end{pmatrix} \\ &\quad + h \overbrace{\begin{pmatrix} \psi_1(1) & \psi_2(1) \\ 0 & 0 \end{pmatrix}}^B \begin{pmatrix} f(x_n + c_1 h, v_1) \\ f(x_n + h, v_2) \end{pmatrix} \\ \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} &= \underbrace{\begin{pmatrix} 1 - \varphi(c_1) & \varphi(c_1) \\ 1 - \varphi(1) & \varphi(1) \end{pmatrix}}_{\tilde{A}} \begin{pmatrix} y_n \\ y_{n-1} \end{pmatrix} \\ &\quad + h \underbrace{\begin{pmatrix} \psi_1(c_1) & \psi_2(c_1) \\ \psi_1(1) & \psi_2(1) \end{pmatrix}}_{\tilde{B}} \begin{pmatrix} f(x_n + c_1 h, v_1) \\ f(x_n + h, v_2) \end{pmatrix} \end{aligned} \tag{9.12}$$

where

$$\begin{aligned} \varphi(x) &= -\frac{6}{5+9c_1} \left(\frac{x^3}{3} - \frac{x^2}{2}(1+c_1) + xc_1 \right) \\ \psi_1(x) &= \frac{x(x+1)}{(1-c_1)(5+9c_1)} (5-3x) \\ \psi_2(x) &= \frac{x(x+1)}{(1-c_1)(5+9c_1)} ((2c_1+1)x - c_1(3c_1+2)). \end{aligned}$$

We know from Exercise V.3.7 that the method is A -stable if and only if $c_1 \geq (\sqrt{17}-1)/8$. For the study of its G -stability we assume that after an appropriate

scaling of G , $g_{11} = 1$. By Lemma 9.5ii the matrix G must then be of the form (recall that $\xi_0 = (1, 1)^T$)

$$G = \begin{pmatrix} 1 & \gamma - 1 \\ \gamma - 1 & (\varphi(1) - 1)\gamma + 1 \end{pmatrix}. \quad (9.13)$$

A necessary condition for G to be positive definite is that $\det G > 0$. For $c_1 \geq 0$ this is equivalent to

$$0 < \gamma < \frac{6(1 + c_1)}{5 + 9c_1}. \quad (9.14)$$

Next we use Lemma 9.5i which implies that

$$d_1 = \gamma\psi_1(1), \quad d_2 = \gamma\psi_2(1). \quad (9.15)$$

Inserting (9.13) and (9.15) into the matrix M of (9.6) yields for its lower right block

$$\begin{pmatrix} \psi_1(1) & 0 \\ 0 & \psi_2(1) \end{pmatrix} \begin{pmatrix} 2\gamma\chi_1 - 1 & (\chi_2 + 1)\gamma - 1 \\ (\chi_2 + 1)\gamma - 1 & 2\gamma - 1 \end{pmatrix} \begin{pmatrix} \psi_1(1) & 0 \\ 0 & \psi_2(1) \end{pmatrix} \quad (9.16)$$

where

$$\chi_1 = \frac{\psi_1(c_1)}{\psi_1(1)} = \frac{1}{4}c_1(c_1 + 1)(5 - 3c_1), \quad \chi_2 = \frac{\psi_2(c_1)}{\psi_2(1)} = \frac{c_1^2(c_1 + 1)^2}{2(3c_1^2 - 1)}.$$

A direct computation (see Exercise 2) shows that this 2×2 matrix can not be non-negative definite for $c_1 \geq (\sqrt{17} - 1)/8$ and γ satisfying (9.14). Consequently the considered methods are never G -stable.

In the next subsections we shall show how high-order algebraically stable general linear methods can be constructed.

Multistep Runge-Kutta Methods

An interesting extension of multistep collocation methods are the so-called multistep Runge-Kutta methods. They are defined by the formulas

$$\begin{aligned} y_{n+1} &= \sum_{j=1}^k \alpha_j y_{n+1-j} + h \sum_{j=1}^s b_j f(x_n + c_j h, v_j^{(n)}) \\ v_i^{(n)} &= \sum_{j=1}^k \tilde{a}_{ij} y_{n+1-j} + h \sum_{j=1}^s \tilde{b}_{ij} f(x_n + c_j h, v_j^{(n)}). \end{aligned} \quad (9.17)$$

They obviously form a subclass of the general linear methods (9.2). This is seen by putting $u_n = (y_n, y_{n-1}, \dots, y_{n-k+1})^T$ so that the exact value function is

$$z(x, h) = (y(x), y(x-h), \dots, y(x - (k-1)h))^T.$$

Further, the matrices A and B have the special form

$$A = \begin{pmatrix} \alpha_1 & \cdots & \cdots & \alpha_k \\ 1 & & & 0 \\ & \ddots & & \vdots \\ & & 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} b_1 & \cdots & b_s \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}. \quad (9.18)$$

The order conditions for such methods were derived in Theorem III.8.14. It follows from this theorem that the method (9.17) is of order p , iff

$$1 = \sum_{j=1}^k \alpha_j (1-j)^{\varrho(t)} + \sum_{j=1}^s b_j \mathbf{v}'_j(t) \quad \text{for } t \in T, \quad \varrho(t) \leq p. \quad (9.19)$$

The values $\mathbf{v}'_j(t)$ are given recursively by

$$\mathbf{v}_i(t) = \sum_{j=1}^k \tilde{a}_{ij} (1-j)^{\varrho(t)} + \sum_{j=1}^s \tilde{b}_{ij} \mathbf{v}'_j(t). \quad (9.20)$$

Recall from Corollary II.12.7 that

$$\begin{aligned} \mathbf{v}'_j(\emptyset) &= 0, & \mathbf{v}'_j(\tau) &= 1 \\ \mathbf{v}'_j(t) &= \varrho(t) \mathbf{v}_j(t_1) \cdots \mathbf{v}_j(t_m) & \text{if } t = [t_1, \dots, t_m]. \end{aligned} \quad (9.21)$$

The order conditions (9.19) constitute a system of nonlinear equations in the coefficients of the method. Without any preparation, solving them may be difficult. We therefore introduce additional assumptions which simplify the construction of multistep Runge-Kutta methods.

Simplifying Assumptions

The conditions $B(p)$, $C(\eta)$ and $D(\xi)$ of Sect. IV.5 were useful for the construction of high-order implicit Runge-Kutta methods. Burrage (1988) showed how these simplifying assumptions can be extended to general linear methods. In the sequel we specialize his approach to multistep Runge-Kutta methods. We consider the assumptions

$$\begin{aligned} B(p) : & \quad q \sum_{j=1}^s b_j c_j^{q-1} + \sum_{j=1}^k \alpha_j (1-j)^q = 1 & q = 1, \dots, p; \\ C(\eta) : & \quad q \sum_{j=1}^s \tilde{b}_{ij} c_j^{q-1} + \sum_{j=1}^k \tilde{a}_{ij} (1-j)^q = c_i^q & q = 1, \dots, \eta, \text{ all } i; \\ D_A(\xi) : & \quad q \sum_{i=1}^s b_i c_i^{q-1} \tilde{a}_{ij} = \alpha_j (1 - (1-j)^q) & q = 1, \dots, \xi, \text{ all } j; \\ D_B(\xi) : & \quad q \sum_{i=1}^s b_i c_i^{q-1} \tilde{b}_{ij} = b_j (1 - c_j^q) & q = 1, \dots, \xi, \text{ all } j. \end{aligned}$$

Condition $B(p)$ is equivalent to the order conditions (9.19) for bushy trees. Condition $C(\eta)$ means that $\mathbf{v}_j(t)$, defined by (9.20), satisfies

$$\mathbf{v}_j(t) = c_j^{\varrho(t)} \quad \text{for } \varrho(t) \leq \eta. \quad (9.22)$$

We remark that the preconsistency condition (9.7) with $\xi_0 = (1, \dots, 1)^T$,

$$\sum_{j=1}^k \alpha_j = 1, \quad \sum_{j=1}^k \tilde{a}_{ij} = 1 \quad \text{for } i = 1, \dots, s, \quad (9.23)$$

is obtained by putting $q = 0$ in $B(p)$ and $C(\eta)$. The condition $D(\xi)$ for Runge-Kutta methods splits into $D_A(\xi)$ and $D_B(\xi)$. However, under certain assumptions one of these conditions is automatically satisfied.

Lemma 9.9. *Suppose that the coefficients c_1, \dots, c_s of a multistep Runge-Kutta method are distinct and $b_i \neq 0$. Then,*

- i) $B(\xi + k - 1), C(k - 1), D_B(\xi) \implies D_A(\xi),$
- ii) $B(\xi + s), C(s), D_A(\xi) \implies D_B(\xi),$
- iii) $B(\eta + s), D_A(s), D_B(s) \implies C(\eta).$

Proof. The first two implications are a consequence of the identity

$$\begin{aligned} & \sum_{j=1}^k \left(q \sum_{i=1}^s b_i c_i^{q-1} \tilde{a}_{ij} - \alpha_j (1 - (1-j)^q) \right) (1-j)^\ell \\ &= -\ell \sum_{j=1}^s \left(q \sum_{i=1}^s b_i c_i^{q-1} \tilde{b}_{ij} - b_j (1 - c_j^q) \right) c_j^{\ell-1} \end{aligned}$$

which holds under the assumptions $C(\ell)$ and $B(q + \ell)$. The last implication can be proved similarly. \square

The fundamental theorem, which generalizes Theorem IV.5.1, is

Theorem 9.10 (Burrage 1988). *If the coefficients of a multistep Runge-Kutta method (9.17) satisfy the simplifying assumptions $B(p)$, $C(\eta)$, $D_A(\xi)$, $D_B(\xi)$ with $p \leq \eta + \xi + 1$ and $p \leq 2\eta + 2$, then the method is of order p .*

Proof. The conditions $C(\eta)$ and $D_A(\xi)$, $D_B(\xi)$ allow the reduction of order conditions of trees as sketched in Fig. 7.1 and Fig. 7.2 of Sect. II.7, respectively. Under the restrictions $p \leq \eta + \xi + 1$ and $p \leq 2\eta + 2$ all order conditions reduce to those for bushy trees which are satisfied by $B(p)$. \square

Remember that we are searching for high-order algebraically stable methods. Due to the Daniel-Moore conjecture (Theorem V.4.4) the order is restricted by $p \leq 2s$. It is therefore natural to look for methods satisfying $B(2s)$, $C(s)$ and

$D_A(s)$, $D_B(s)$. They will be of order $2s$ by Theorem 9.10 and are an extension of the Runge-Kutta methods based on Gauss quadrature. Let us begin by studying the condition $B(2s)$.

Quadrature Formulas

Because of (9.23) condition $B(p)$ of the preceding subsection is equivalent to

$$\sum_{j=1}^s b_j f(c_j) = \sum_{j=1}^k \alpha_j \int_{1-j}^1 f(x) dx, \quad \deg f \leq p-1, \quad (9.24)$$

where f stands for a polynomial of degree at most $p-1$. For the construction of such quadrature formulas it is useful to consider the bilinear form

$$\langle f, g \rangle = \sum_{j=1}^k \alpha_j \int_{1-j}^1 f(x) g(x) dx = \int_{1-k}^1 \omega(x) f(x) g(x) dx, \quad (9.25)$$

where $\omega(x)$ is the step-function sketched in Fig. 9.1. Under the assumption

$$\alpha_k \geq 0, \quad \alpha_k + \alpha_{k-1} \geq 0, \dots, \quad \alpha_k + \dots + \alpha_2 \geq 0, \quad \alpha_k + \dots + \alpha_1 = 1, \quad (9.26)$$

$\omega(x)$ is non-negative and (9.25) becomes an inner product on the space of real polynomials. We call the quadrature formula (9.24) *interpolatory* if $B(s)$ holds. This implies that

$$b_i = \int_{1-k}^1 \omega(x) \ell_i(x) dx, \quad \ell_i(x) = \prod_{\substack{l=1 \\ l \neq i}}^s \frac{(x - c_l)}{(c_i - c_l)}. \quad (9.27)$$

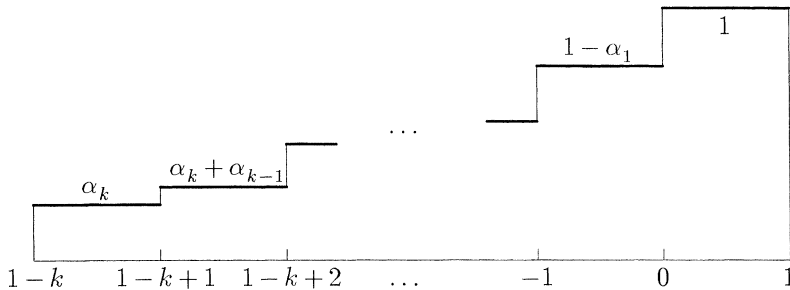


Fig. 9.1. Weight function for the inner product (9.25)

The following results on Gaussian quadrature and orthogonal polynomials are classical.

Lemma 9.11. Let $M(x) = (x - c_1) \cdots (x - c_s)$. An interpolatory quadrature formula satisfies $B(s + m)$ if and only if

$$\sum_{j=1}^k \alpha_j \int_{1-j}^1 M(x) x^{q-1} dx = 0 \quad \text{for } q = 1, \dots, m. \quad \square$$

Let $p_s(x)$ be the polynomial of degree s which is orthogonal with respect to (9.25) to all polynomials of degree $s - 1$. Lemma 9.11 then states that a quadrature formula (9.24) is of order $2s$ iff $M(x)$ is a scalar multiple of $p_s(x)$. The polynomials $p_s(x)$ which depend on $\alpha_1, \dots, \alpha_k$ via the bilinear form (9.25) can be computed from a standard three term recursion

$$\begin{aligned} p_0(x) &= 1, & p_1(x) &= x - \beta_0 \\ p_{s+1}(x) &= (x - \beta_s)p_s(x) - \gamma_s p_{s-1}(x) \end{aligned} \quad (9.28)$$

where

$$\beta_s = \frac{\langle xp_s, p_s \rangle}{\langle p_s, p_s \rangle}, \quad \gamma_s = \frac{\langle p_s, p_s \rangle}{\langle p_{s-1}, p_{s-1} \rangle}. \quad (9.29)$$

Obviously this is only possible if $\langle p_j, p_j \rangle \neq 0$ for $j = 1, \dots, s$. This is certainly the case under the assumption (9.26).

Lemma 9.12. If $\alpha_1, \dots, \alpha_k$ satisfy (9.26) then all zeros of $p_s(x)$ are real, simple and lie in the open interval $(1 - k, 1)$. \square

For the construction of algebraically stable methods, quadrature formulas with positive weights will be of particular interest. Sufficient conditions for this property are given in the following theorem.

Theorem 9.13. If the quadrature formula (9.24) is of order $p \geq 2s - 1$ and if $\alpha_1, \dots, \alpha_k$ satisfy (9.26), then

$$b_i > 0 \quad \text{for } i = 1, \dots, s. \quad \square$$

Algebraically Stable Methods of Order $2s$

... the analysis of the algebraic stability properties of multivalued methods ... is not as difficult as was generally thought ...
(Burrage 1987)

Following Burrage (1987) we consider the following class of multistep Runge-Kutta methods.

Definition 9.14. Let $\alpha_1, \dots, \alpha_k$ with $\sum \alpha_j = 1$ and $\alpha_k \neq 0$ be given such that the zeros c_1, \dots, c_s of $p_s(x)$ (Formula (9.28)) are real and simple. We then denote

by $E(\alpha_1, \dots, \alpha_k)$ the multistep Runge-Kutta method (9.17) whose coefficients are given by

$$\begin{aligned} b_i &= \sum_{j=1}^k \alpha_j \int_{1-j}^1 \ell_i(x) dx, & i = 1, \dots, s, \\ \tilde{a}_{ij} &= \frac{\alpha_j}{b_j} \int_{1-j}^1 \ell_i(x) dx, & i = 1, \dots, s; j = 1, \dots, k \\ \tilde{b}_{ij} &= \frac{b_j}{b_i} \int_{c_j}^1 \ell_i(x) dx, & i = 1, \dots, s; j = 1, \dots, s \end{aligned}$$

where $\ell_i(x)$ is the function of (9.27).

The definitions of c_i and b_i imply $B(2s)$ by Lemma 9.11. The formulas for \tilde{a}_{ij} and \tilde{b}_{ij} are equivalent to $D_A(s)$ and $D_B(s)$, respectively. Lemma 9.9iii thus implies $C(s)$ and Theorem 9.10 finally proves that the considered methods are of order $2s$. The following theorem gives sufficient conditions for the algebraic stability of these methods.

Theorem 9.15 (Burrage 1987). *If $\alpha_j \geq 0$ for $j = 1, \dots, k$, then the method $E(\alpha_1, \dots, \alpha_k)$ is G -stable with*

$$G = \text{diag}(1, \alpha_2 + \dots + \alpha_k, \dots, \alpha_{k-1} + \alpha_k, \alpha_k). \quad (9.30)$$

Proof. For multistep Runge-Kutta methods the preconsistency vector is given by $\xi_0 = (1, 1, \dots, 1)^T$. With the matrix G of (9.30) it therefore follows from Lemma 9.5 that

$$d_i = b_i \quad \text{for} \quad i = 1, \dots, s. \quad (9.31)$$

By Theorem 9.13 this implies $d_i > 0$ so that the first condition for algebraic stability is satisfied. In order to verify that the matrix M of (9.6) is non-negative definite, we transform it by a suitable matrix. We put

$$V = \left(c_i^{j-1} \right)_{i,j=1,\dots,s} \quad \text{and} \quad \alpha = (\alpha_1, \dots, \alpha_k)^T. \quad (9.32)$$

A straightforward calculation using the simplifying assumptions $D_A(s)$, $D_B(s)$ and $B(2s)$ shows that

$$\begin{pmatrix} I & 0 \\ 0 & V^T \end{pmatrix} M \begin{pmatrix} I & 0 \\ 0 & V \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & W^T \end{pmatrix} \widehat{M} \begin{pmatrix} I & 0 \\ 0 & W \end{pmatrix} \quad (9.33)$$

where

$$W = \left(\frac{1}{j} (1-i)^j \right)_{\substack{i=1,\dots,k \\ j=1,\dots,s}}$$

and the $2k \times 2k$ matrix \widehat{M} is given by

$$\widehat{M} = \begin{pmatrix} Z & Z \\ Z & Z \end{pmatrix}, \quad Z = \text{diag}(\alpha_1, \dots, \alpha_k) - \alpha \alpha^T. \quad (9.34)$$

Since $\alpha_j \geq 0$ and $\sum \alpha_j = 1$ it follows from the Cauchy-Schwarz inequality that

$$x^T Z x = \sum_{j=1}^k \alpha_j x_j^2 - \left(\sum_{j=1}^k \alpha_j x_j \right)^2 \geq 0$$

Therefore the matrix Z , and hence also \widehat{M} , are non-negative definite matrices. This completes the proof of the theorem. \square

One can ask what are the advantages of the methods $E(\alpha_1, \dots, \alpha_k)$ with $k > 1$ over the s -stage Gauss Runge-Kutta methods of order $2s$. All these methods have the same order and are algebraically stable for $\alpha_j \geq 0$.

- The Gauss methods have a stability function whose value at infinity satisfies $|R(\infty)| = 1$. In contrast, the new methods allow the spectral radius $\varrho(S(\infty))$ to be smaller than 1, which improves stability at infinity. For example, numerical investigations of the case $s = 2$, $k = 2$ show that $\varrho(S(\infty))$ has the minimal value $\sqrt{2} - 1 \approx 0.41421$ for $\alpha_1 = 12\sqrt{2} - 16$ and $\alpha_2 = 1 - \alpha_1$ (see Exercise 7). There are some indications that L -stable methods do not exist: if we could find methods with an internal stage, say $v_s^{(n)}$, equal to y_{n+1} , then the method would be L -stable. Unfortunately, this would imply $c_s = 1$, which is in contradiction to Lemma 9.12 and to $\alpha_j \geq 0$.

- The eigenvalues of the Runge-Kutta matrix of the Gauss methods are complex (with the exception of one real eigenvalue, if s is odd). Can we hope that, for a suitable choice of $\alpha_j \geq 0$, all eigenvalues of \tilde{B} become real? Numerical computations for $s = 2$ and $k = 2$ indicate that this is not possible.

B-Convergence

Many results of Sections IV.14 and IV.15 have a straightforward extension to general linear methods. The following theorem corresponds to Theorems IV.14.2, IV.14.3, and IV.14.4 and is proved in the same way:

Theorem 9.16. *Let f be continuously differentiable and satisfy (9.1). If the matrix \tilde{B} of method (9.2) is invertible and if*

$$h\nu < \alpha_0(\tilde{B}^{-1}),$$

then the nonlinear system (9.2b) has a unique solution. \square

The next results give estimates of the local and global errors. We formulate these results only for multistep Runge-Kutta methods, because in this case the definitions of $C(\eta)$ and $B(p)$ are already available. In analogy to Runge-Kutta

methods we say that method (9.17) has *stage order* q , if $C(q)$ and $B(q)$ are satisfied. Recall that for the definition of the local error

$$\delta_h(x) = y_1 - y(x+h)$$

one assumes that $y_i = y(x+ih)$ for $i = 1-k, \dots, 0$ lie on the exact solution.

Theorem 9.17. *Suppose that the differential equation satisfies (9.1). If the matrix \tilde{B} is invertible, if $\alpha_0(\tilde{B}^{-1}) > 0$ and if the stage order is q , then the local error of method (9.17) satisfies*

$$\|\delta_h(x)\| \leq Ch^{q+1} \max_{\xi \in [x-(k-1)h, x+h]} \|y^{(q+1)}(\xi)\| \quad \text{for } h\nu \leq \alpha < \alpha_0(\tilde{B}^{-1})$$

where C depends only on the coefficients of the method and on α . \square

This result, which corresponds to Proposition IV.15.1, is of particular interest for multistep collocation methods, for which the stage order $q = s + k - 1$ is maximal. The global error allows the following estimate, which extends Theorem IV.15.3.

Theorem 9.18. *Suppose, in addition to the assumptions of Theorem 9.17, that the method (9.17) is algebraically stable.*

a) *If $\nu > 0$ then the global error satisfies for $h\nu \leq \alpha < \alpha_0(\tilde{B}^{-1})$*

$$\|y_n - y(x_n)\| \leq h^q \frac{e^{C_1\nu(x_n-x_0)} - 1}{C_1\nu} C_2 \max_{x \in [x_0, x_n]} \|y^{(q+1)}(x)\|.$$

b) *If $\nu \leq 0$ then (for all $h > 0$)*

$$\|y_n - y(x_n)\| \leq h^q (x_n - x_0) C_2 \max_{x \in [x_0, x_n]} \|y^{(q+1)}(x)\|.$$

The constants C_1 and C_2 depend only on the coefficients of the method and (for case a) on α . \square

In contrast to the results of Sect. IV.15 the above theorem holds only for a constant step size implementation.

Exercises

1. Show that for Runge-Kutta methods, where $A = (1)$, $\tilde{A} = \mathbb{1}$, both definitions of algebraic stability (IV.12.5 and V.9.3) are the same.
2. Prove in detail the statement of Example 9.8, that the 2-step 2-stage collocation methods with $c_2 = 1$ (and $c_1 \neq 1$) are not G -stable.
Hint. The non-negativity of the matrix (9.16) implies $\gamma \geq 1/2$ and by considering its determinant,

$$\gamma(4\chi_1 - (1 + \chi_2)^2) \geq 2(\chi_1 - \chi_2).$$

This inequality contradicts (9.14).

3. If a multistep Runge-Kutta method with distinct c_i and $c_i \geq 0$ satisfies the assumptions $B(s + k + \xi)$ and $C(s + k - 1)$, then it also satisfies $D_B(\xi)$.

Hint. Show that

$$\sum_{j=1}^k \left(q \sum_{i=1}^s b_i c_i^{q-1} \tilde{a}_{ij} - \alpha_j (1 - (1-j)^q) \right) (r(1) - r(1-j)) = 0$$

for all polynomials $r(x)$ of degree $\leq s + k - 1$ which satisfy $r(c_1) = \dots = r(c_s) = 0$. For given j , construct such a polynomial which also satisfies

$$r(1-j) = 1, \quad r(1-i) = 0 \quad \text{for } i = 1, \dots, k \quad \text{and} \quad i \neq j.$$

4. Disprove the conjecture of Burrage (1988) that for every k and s there exist zero-stable multistep Runge-Kutta methods of order $2s + k - 1$.

Hint. Consider the case $s = 1$ so that these methods are equivalent to one-leg methods and consult a result of Dahlquist (1983).

5. (Burrage 1988). Show that there exists a zero-stable multistep Runge-Kutta method with $s = 2$ and $k = 2$ which is of order 5.

Result. $c_{1,2} = (\sqrt{7} \pm \sqrt{2})/5$

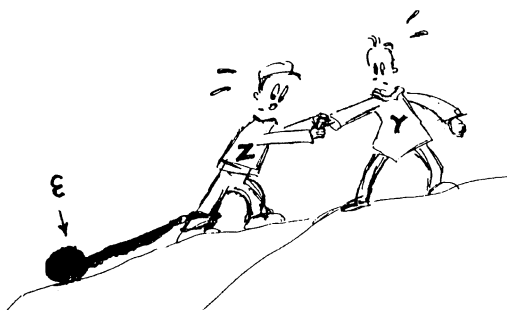
6. (Stability at infinity). If a multistep Runge-Kutta method satisfies $D_A(s)$ and $D_B(s)$ then we have, e.g., for $s = 2$ and $k = 2$,

$$S(\infty) = \begin{pmatrix} \alpha_1 & \alpha_2 \\ 1 & 0 \end{pmatrix} - \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 - c_1 & 1 - c_2 \\ 1 - c_1^2 & 1 - c_2^2 \end{pmatrix}^{-1} \begin{pmatrix} \alpha_1 & 2\alpha_2 \\ \alpha_1 & 0 \end{pmatrix}.$$

Formulate this result also for general s and k .

7. Verify that for the method $E(\alpha_1, \alpha_2)$ with $0 \leq \alpha_1 \leq 1$, $\alpha_2 = 1 - \alpha_1$, the spectral radius $\rho(S(\infty))$ is minimal for $\alpha_1 = 12\sqrt{2} - 16$.

Chapter VI. Singular Perturbation Problems and Index 1 Problems



(Drawing by G. Di Marzo)

Singular perturbation problems (SPP) form a special class of problems containing a parameter ε . When this parameter is small, the corresponding differential equation is stiff; when ε tends to zero, the differential equation becomes differential algebraic. This chapter investigates the numerical solution of such singular perturbation problems. This allows us to understand many phenomena observed for very stiff problems. Much insight is obtained by studying the limit case $\varepsilon = 0$ (“the reduced system” or “problem of index 1”) which is usually much easier to analyze.

We start by considering the limit case $\varepsilon = 0$. Two numerical approaches – the ε -embedding method and the state space form method – are investigated in Sect. VI.1. We then analyze multistep methods in Sect. VI.2, Runge-Kutta methods in Sect. VI.3, Rosenbrock methods in Sect. VI.4 and extrapolation methods in Sect. VI.5. Convergence is studied for singular perturbation problems and for semi-explicit differential-algebraic systems of “index 1”.

VI.1 Solving Index 1 Problems

Singular perturbation problems (SPP) have several origins in applied mathematics. One comes from fluid dynamics and results in linear boundary value problems containing a small parameter ε (the coefficient of viscosity) such that for $\varepsilon \rightarrow 0$ the differential equation loses the highest derivative (see Exercise 1 below). Others originate in the study of nonlinear oscillations with *large* parameters (van der Pol 1926, Dorodnicyn 1947) or in the study of chemical kinetics with slow and fast reactions (see e.g., Example (IV.1.4)).

Asymptotic Solution of van der Pol's Equation

The classical paper of Dorodnicyn (1947) studied the van der Pol Equation (IV.1.5') with large μ , i.e., with small ε . The investigation becomes a little easier if we use Liénard's coordinates (see Exercise I.16.8). In Eq. (IV.1.5'), written here as

$$\varepsilon z'' + (z^2 - 1)z' + z = 0, \quad (1.1)$$

we insert the identity

$$\varepsilon z'' + (z^2 - 1)z' = \frac{d}{dx} \underbrace{\left(\varepsilon z' + \left(\frac{z^3}{3} - z \right) \right)}_{:= y}$$

so that (1.1) becomes

$$\begin{aligned} y' &= -z & &=: f(y, z) \\ \varepsilon z' &= y - \left(\frac{z^3}{3} - z \right) & &=: g(y, z). \end{aligned} \quad (1.2)$$

Fig. 1.1 shows solutions of Eq. (1.2) with $\varepsilon = 0.03$ in the (y, z) -plane. One observes rapid movements towards the manifold M defined by $y = z^3/3 - z$, close to which the solution becomes smooth. In order to approximate the solution for very small ε , we set $\varepsilon = 0$ in (1.2) and obtain the so-called *reduced* system

$$\begin{aligned} y' &= -z & &= f(y, z) \\ 0 &= y - \left(\frac{z^3}{3} - z \right) & &= g(y, z). \end{aligned} \quad (1.2')$$

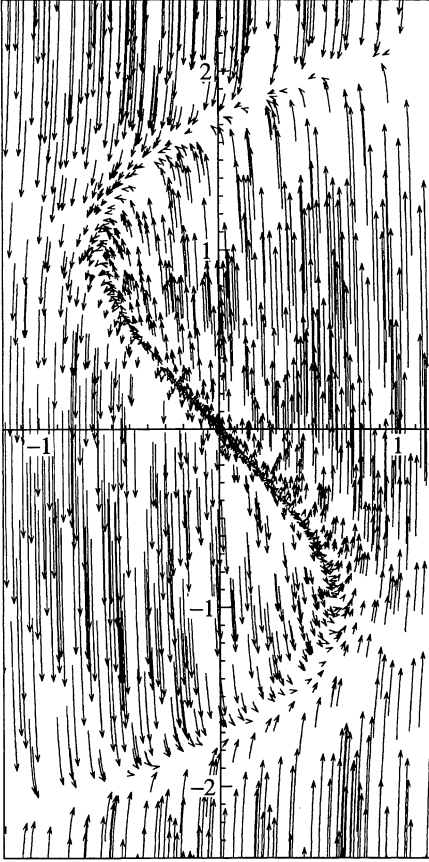


Fig. 1.1. Solutions of SPP (1.2)

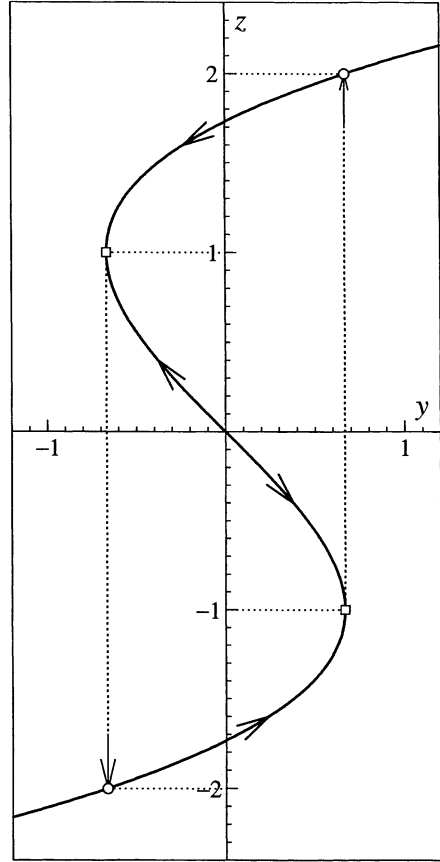


Fig. 1.2. Reduced problem (1.2')

While (1.2) has no analytic solution, (1.2') can easily be solved to give

$$y' = -z = (z^2 - 1)z' \quad \text{or} \quad \ln|z| - \frac{z^2}{2} = x + C. \quad (1.3)$$

Equation (1.2') is called a *differential algebraic equation* (DAE), since it combines a differential equation (first line) with an algebraic equation (second line). Such a problem only makes sense if the initial values are *consistent*, i.e., lie on the manifold M . The points of M with coordinates $y = \pm 2/3$, $z = \mp 1$ are of special interest (Fig. 1.2): at these points the partial derivative $g_z = \partial g / \partial z$ vanishes and the defining manifold is no longer “transversal” to the direction of the fast movement. Here the solutions of (1.2') cease to exist, while the solutions of the full problem (1.2) for $\varepsilon \rightarrow 0$ jump with “infinite” speed to the opposite manifold. For $-1 < z < 1$ the manifold M is *unstable* for the solution of (1.2) (here $g_z > 0$), otherwise M is *stable* ($g_z < 0$).

We demonstrate the power of the reduced equation by answering the question:

what is the period T of the limit cycle solution of van der Pol's equation for $\varepsilon \rightarrow 0$? Fig. 1.2 shows that the asymptotic value of T is just twice the time which $z(x)$ of (1.3) needs to advance from $z = -2$ to $z = -1$, i.e.,

$$T = 3 - 2 \ln 2. \quad (1.4)$$

This is the first term of Dorodnicyn's asymptotic formula. We also see that $z(x)$ reaches its largest values (i.e., crosses the Poincaré cut $z' = 0$, see Fig. I.16.2) at $z = \pm 2$. We thus have the curious result that the limit cycle of van der Pol's equation (1.1) has the same asymptotic initial value $z = 2$ and $z' = 0$ for $\varepsilon \rightarrow 0$ and for $\varepsilon \rightarrow \infty$ (see Eq. (I.16.10)).

The ε -Embedding Method for Problems of Index 1

We now want to study the behaviour of the *numerical solution* for $\varepsilon \rightarrow 0$. This will give us insight into many phenomena encountered for very stiff equations and also suggest advantageous numerical procedures for stiff and differential-algebraic equations. Let an arbitrary singular perturbation problem be given,

$$y' = f(y, z) \quad (1.5a)$$

$$\varepsilon z' = g(y, z), \quad (1.5b)$$

where y and z are vectors; suppose that f and g are sufficiently often differentiable vector functions of the same dimensions as y and z , respectively. The corresponding *reduced* equation is the DAE

$$y' = f(y, z) \quad (1.6a)$$

$$0 = g(y, z), \quad (1.6b)$$

whose initial values are *consistent* if $0 = g(y_0, z_0)$. A general assumption of the present chapter will be that the Jacobian

$$g_z(y, z) \quad \text{is invertible} \quad (1.7)$$

in a neighbourhood of the solution of (1.6). Equation (1.6b) then possesses a locally unique solution $z = G(y)$ ("Implicit Function Theorem") which inserted into (1.6a) gives

$$y' = f(y, G(y)), \quad (1.8)$$

the so-called "state space form", an ordinary differential system. Under the assumption (1.7), Eq. (1.6) is said to be a differential-algebraic equation of *index 1*.

An interesting approach for solving (1.6) is to apply some numerical method to the SPP (1.5) and to put $\varepsilon = 0$ in the resulting formulas. Let us illustrate this approach for Runge-Kutta methods. Applied to the system (1.5) we obtain

$$Y_{ni} = y_n + h \sum_{j=1}^s a_{ij} f(Y_{nj}, Z_{nj}) \quad (1.9a)$$

$$\varepsilon Z_{ni} = \varepsilon z_n + h \sum_{j=1}^s a_{ij} g(Y_{nj}, Z_{nj}) \quad (1.9b)$$

$$y_{n+1} = y_n + h \sum_{i=1}^s b_i f(Y_{ni}, Z_{ni}) \quad (1.9c)$$

$$\varepsilon z_{n+1} = \varepsilon z_n + h \sum_{i=1}^s b_i g(Y_{ni}, Z_{ni}). \quad (1.9d)$$

We now suppose that the RK matrix (a_{ij}) is invertible and obtain from (1.9b)

$$hg(Y_{ni}, Z_{ni}) = \varepsilon \sum_{j=1}^s \omega_{ij} (Z_{nj} - z_n), \quad (1.10)$$

where the ω_{ij} are the elements of the inverse of (a_{ij}) . Inserting this into (1.9d) makes the definition of z_{n+1} independent of ε . We thus put without more ado $\varepsilon = 0$ and obtain

$$Y_{ni} = y_n + h \sum_{j=1}^s a_{ij} f(Y_{nj}, Z_{nj}) \quad (1.11a)$$

$$0 = g(Y_{ni}, Z_{ni}) \quad (1.11b)$$

$$y_{n+1} = y_n + h \sum_{i=1}^s b_i f(Y_{ni}, Z_{ni}) \quad (1.11c)$$

$$z_{n+1} = \left(1 - \sum_{i,j=1}^s b_i \omega_{ij}\right) z_n + \sum_{i,j=1}^s b_i \omega_{ij} Z_{nj}. \quad (1.11d)$$

Here

$$1 - \sum_{i,j=1}^s b_i \omega_{ij} = R(\infty) \quad (1.11e)$$

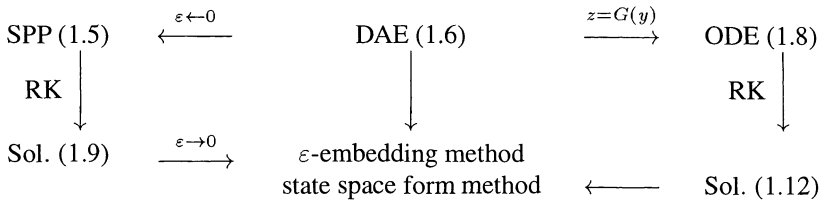
(see Eq. (IV.3.15)), where $R(z)$ is the stability function of the method.

State Space Form Method

The numerical solution (y_{n+1}, z_{n+1}) of the above approach will usually *not* lie on the manifold $g(y, z) = 0$. However, this can easily be repaired by replacing (1.11d) by the condition

$$0 = g(y_{n+1}, z_{n+1}). \quad (1.12)$$

Then, we do not only have $Z_{nj} = G(Y_{nj})$ (see (1.11b)), but also $z_{n+1} = G(y_{n+1})$. In this case the method (1.11a–c), (1.12) is *identical* to the solution of the state space form (1.8) with the same Runge-Kutta method. This will be called the *state space form method*. The whole situation is summarized in the following diagram:



Of special importance here are *stiffly accurate* methods, i.e., methods which satisfy

$$a_{si} = b_i \quad \text{for } i = 1, \dots, s. \quad (1.13)$$

This means that $y_{n+1} = Y_{ns}$, $z_{n+1} = Z_{ns}$ and (1.12) is satisfied anyway. Hence for stiffly accurate methods the ε -embedding method and the state space form method are identical. For this reason, Griepentrog & März (1986) denote such methods IRK(DAE).

Both approaches have their own merits. Theoretical results for the ε -embedding method yield insight into the method when applied to singular perturbation problems. Moreover, this approach can easily be extended to more general situations, where the algebraic relation is not explicitly separated from the differential equation (see below). The state space form method, on the other hand, has the advantage that it is not restricted to implicit methods. Applying an explicit Runge-Kutta method or a multistep method to Eq. (1.8) is certainly a method of choice for semi-explicit index 1 equations. No new theory is necessary in this case.

A Transistor Amplifier

... auf eine merkwürdige Tatsache aufmerksam machen, das ist die außerordentlich grosse Zahl berühmter Mathematiker, die aus Königsberg stammen ... : Kant 1724, Richelot 1808, Hesse 1811, Kirchhoff 1824, Carl Neumann 1832, Clebsch 1833, Hilbert 1862.
(F. Klein, Entw. der Math., p. 159)

Very often, differential-algebraic problems arising in practice are not at once in the semi-explicit form (1.6), but rather in the form $Mu' = \varphi(u)$ where M is a constant *singular* matrix.

As an example we compute the amplifier of Fig. 1.3, where $U_e(t)$ is the entry voltage, $U_b = 6$ the operating voltage, $U_i(t)$ ($i = 1, 2, 3, 4, 5$) the voltages at the nodes 1, 2, 3, 4, 5, and $U_5(t)$ the output voltage. The current through a resistor satisfies $I = U/R$ (Ohm 1827), the current through a capacitor $I = C \cdot dU/dt$, where R and C are constants and U the voltage. The transistor acts as amplifier in that the current from node 4 to node 3 is 99 times larger than that from node 2 to node 3 and depends on the voltage difference $U_3 - U_2$ in a nonlinear way. Kirchhoff's law (a Königsberg discovery) says that the sum of currents entering a node vanishes. This law applied to the 5 nodes of Fig. 1.3 leads to the following equations:

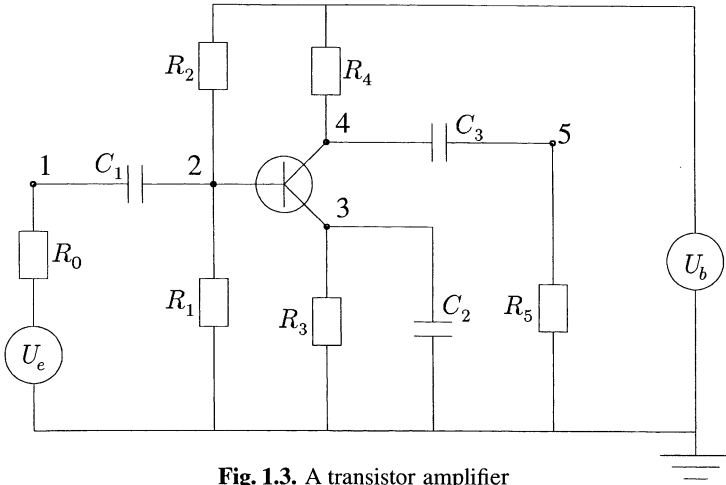


Fig. 1.3. A transistor amplifier

$$\begin{aligned}
 \text{node 1: } & \frac{U_e(t)}{R_0} - \frac{U_1}{R_0} + C_1(U'_2 - U'_1) = 0 \\
 \text{node 2: } & \frac{U_b}{R_2} - U_2 \left(\frac{1}{R_1} + \frac{1}{R_2} \right) + C_1(U'_1 - U'_2) - 0.01 f(U_2 - U_3) = 0 \\
 \text{node 3: } & f(U_2 - U_3) - \frac{U_3}{R_3} - C_2 U'_3 = 0 \\
 \text{node 4: } & \frac{U_b}{R_4} - \frac{U_4}{R_4} + C_3(U'_5 - U'_4) - 0.99 f(U_2 - U_3) = 0 \\
 \text{node 5: } & -\frac{U_5}{R_5} + C_3(U'_4 - U'_5) = 0.
 \end{aligned} \tag{1.14}$$

As constants we adopt the values reported (for a similar problem) by Rentrop, Roche & Steinebach (1989)

$$\begin{aligned}
 f(U) &= 10^{-6} \left(\exp\left(\frac{U}{0.026}\right) - 1 \right) \\
 R_0 &= 1000, \quad R_1 = \dots = R_5 = 9000 \\
 C_k &= k \cdot 10^{-6}, \quad k = 1, 2, 3,
 \end{aligned}$$

and the initial signal is chosen as

$$U_e(t) = 0.4 \cdot \sin(200\pi t). \tag{1.15}$$

Equations (1.14) are of the form $Mu' = \varphi(u)$ where

$$M = \begin{pmatrix} -C_1 & C_1 & & & \\ C_1 & -C_1 & & & \\ & & -C_2 & & \\ & & & -C_3 & C_3 \\ & & & C_3 & -C_3 \end{pmatrix}$$

is obviously a singular matrix of rank 3. The sum of the first two and of the last two equations leads directly to two algebraic equations. Introducing e.g.,

$$U_1 - U_2 = y_1, \quad U_3 = y_2, \quad U_4 - U_5 = y_3, \quad U_1 = z_1, \quad U_4 = z_2,$$

transforms equations (1.14) to the form (1.6). *Consistent initial values* must thus satisfy $\varphi_1(u) + \varphi_2(u) = 0$ and $\varphi_4(u) + \varphi_5(u) = 0$. If we put $U_2(0) = U_3(0)$, we have $f(U_2(0) - U_3(0)) = 0$. Since $U_e(0) = 0$, we then easily find consistent initial values, e.g., as

$$U_1(0) = 0, \quad U_2(0) = U_3(0) = \frac{U_b R_1}{R_1 + R_2}, \quad U_4(0) = U_b, \quad U_5(0) = 0. \quad (1.16)$$

Problems of the Form $Mu' = \varphi(u)$

Numerical methods for problems of the form

$$Mu' = \varphi(u), \quad (1.17)$$

where M is a constant matrix, can be derived as follows: we assume that M is regular, apply an ODE method to $u' = M^{-1}\varphi(u)$ and multiply the resulting formulas by M . For Runge-Kutta methods we obtain in this way

$$M(U_{ni} - u_n) = h \sum_{j=1}^s a_{ij} \varphi(U_{nj}) \quad (1.18a)$$

$$u_{n+1} = \left(1 - \sum_{i,j=1}^s b_i \omega_{ij}\right) u_n + \sum_{i,j=1}^s b_i \omega_{ij} U_{nj}, \quad (1.18b)$$

where again (ω_{ij}) is the inverse of (a_{ij}) . The second formula was obtained from

$$M(u_{n+1} - u_n) = h \sum_{i=1}^s b_i \varphi(U_{ni}) \quad (1.18c)$$

in exactly the same way as above (see (1.10)).

Formulas (1.18) also make sense formally when M is a *singular* matrix. In this case, problem (1.17) is mathematically equivalent to a semi-explicit system (1.6) and method (1.18) corresponds to method (1.11). This can be seen as follows: we decompose the matrix M (e.g., by Gaussian elimination with total pivoting) as

$$M = S \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} T, \quad (1.19)$$

where S and T are invertible matrices and the dimension of I represents the rank of M . Inserting this into (1.17), multiplying by S^{-1} , and using the transformed variables

$$Tu = \begin{pmatrix} y \\ z \end{pmatrix} \quad (1.20)$$

gives

$$\begin{pmatrix} y' \\ 0 \end{pmatrix} = S^{-1} \varphi \left(T^{-1} \begin{pmatrix} y \\ z \end{pmatrix} \right) = : \begin{pmatrix} f(y, z) \\ g(y, z) \end{pmatrix}, \quad (1.21)$$

a problem of type (1.6). An initial value u_0 is *consistent* if $\varphi(u_0)$ lies in the range of the matrix M .

Similarly, if (1.19) is inserted into (1.18), and the variables

$$TU_{nj} = \begin{pmatrix} Y_{nj} \\ Z_{nj} \end{pmatrix}, \quad Tu_n = \begin{pmatrix} y_n \\ z_n \end{pmatrix} \quad (1.22)$$

are introduced, Eq. (1.18b) (for Z_{n+1}) and Eq. (1.18c) (for Y_{n+1}) lead precisely to equations (1.11). This means that the diagram

$$\begin{array}{ccc} \text{Problem (1.17)} & \xrightarrow{\text{Transf. (1.20)}} & \text{Problem (1.6)} \\ \text{Meth.} \downarrow (1.18) & & \text{Meth.} \downarrow (1.11) \\ \{u_n\} & \xrightarrow{\text{Transf. (1.22)}} & \{y_n\}, \{z_n\} \end{array} \quad (1.23)$$

commutes. An important consequence of this commutativity is that all results for semi-explicit systems (1.6) and the ε -embedding method (1.11) (existence of a numerical solution, convergence, asymptotic expansions, ...) also apply to implicit problems (1.17) with singular M and method (1.18).

All codes, such as RADAU5, which have an option for implicit differential equations (1.17) can thus be applied directly. This has been done for problem (1.14) with initial values (1.16), integration interval $0 \leq x \leq 0.2$, and $Tol = 10^{-4}$. The code computed the solution $U_5(t)$ displayed in Fig. 1.4 in 556 (accepted) steps. The comparison with the entry voltage $U_e(t)$ shows that our amplifier is working. See also Hairer, Lubich & Roche (1989), p. 108-111 for a more elaborate example.

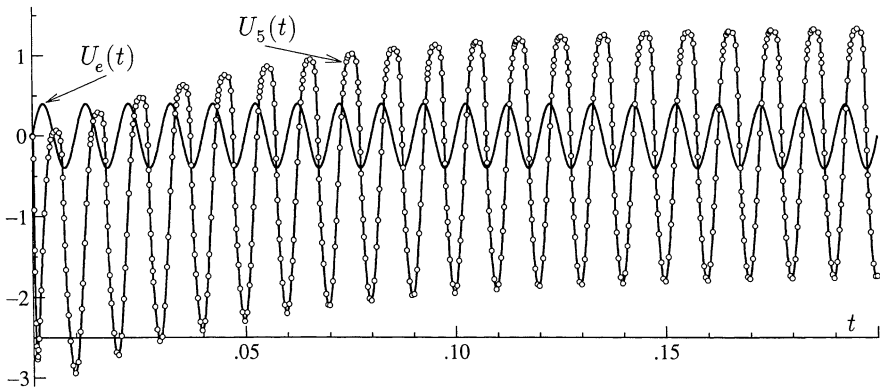


Fig. 1.4. Computed solution of amplifier problem (1.14)

Convergence of Runge-Kutta Methods

If the method is stiffly accurate, the numerical solutions (1.11) are equivalent to those of the *ordinary* equation (1.8). Therefore the convergence of the solutions is described by Theorems II.3.4 and II.3.6 as

$$y_n - y(x_n) = \mathcal{O}(h^p), \quad z_n - z(x_n) = \mathcal{O}(h^p), \quad (1.24)$$

where p is the *classical* order of the method (the second formula follows from a Lipschitz condition for G). For *general* methods, the estimate (1.24) remains valid for y_n , because (1.11a,b,c) are independent of z_n and do not change if (1.11d) is replaced by (1.12). Thus we only have to prove a convergence result for z_n . An essential ingredient of the following theorem is the *stage order* q of the method, i.e., condition $C(q)$ of Sect. II.7 or IV.5.

Theorem 1.1. *Suppose that the system (1.6) satisfies (1.7) in a neighbourhood of the exact solution $(y(x), z(x))$ and assume the initial values are consistent. Consider a Runge-Kutta method of order p , stage order q and with invertible matrix A . Then the numerical solution of (1.11a–d) has global error*

$$z_n - z(x_n) = \mathcal{O}(h^r) \quad \text{for} \quad x_n - x_0 = nh \leq \text{Const}, \quad (1.25)$$

where

- a) $r = p$ for stiffly accurate methods,
- b) $r = \min(p, q + 1)$ if the stability function satisfies $-1 \leq R(\infty) < 1$,
- c) $r = \min(p - 1, q)$ if $R(\infty) = +1$.
- d) If $|R(\infty)| > 1$, the numerical solution diverges.

Proof. Part (a) has already been discussed. For the remaining cases we proceed as follows: we first observe that Condition $C(q)$ and order p imply

$$z(x_n + c_i h) = z(x_n) + h \sum_{j=1}^s a_{ij} z'(x_n + c_j h) + \mathcal{O}(h^{q+1}) \quad (1.26a)$$

$$z(x_{n+1}) = z(x_n) + h \sum_{i=1}^s b_i z'(x_n + c_i h) + \mathcal{O}(h^{p+1}). \quad (1.26b)$$

Since A is invertible we can compute $z'(x_n + c_j h)$ from (1.26a) and insert it into (1.26b). This gives

$$z(x_{n+1}) = \varrho z(x_n) + b^T A^{-1} \hat{Z}_n + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+1}) \quad (1.27)$$

where $\varrho = 1 - b^T A^{-1} \mathbb{1} = R(\infty)$ and $\hat{Z}_n = (z(x_n + c_1 h), \dots, z(x_n + c_s h))^T$. We then denote the global error by $\Delta z_n = z_n - z(x_n)$, and $\Delta Z_n = Z_n - \hat{Z}_n$. Subtracting (1.27) from (1.11d) yields

$$\Delta z_{n+1} = \varrho \Delta z_n + b^T A^{-1} \Delta Z_n + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+1}). \quad (1.28)$$

Our next aim is to estimate ΔZ_n . For this we have to consider the y -component of the system. Due to (1.11a–c) the values y_n, Y_{ni} are those of the Runge-Kutta method applied to (1.8). It thus follows from Theorem II.8.1 that $y_n - y(x_n) = e_p(x_n)h^p + \mathcal{O}(h^{p+1})$. Since Eq. (1.26a) also holds with $z(x)$ replaced by $y(x)$, we can subtract this formula from (1.11a) and so obtain

$$Y_{ni} - y(x_n + c_i h) = y_n - y(x_n) + h \sum_{j=1}^s a_{ij} \left(f(Y_{nj}, G(Y_{nj})) - f(y(x_n + c_j h), G(y(x_n + c_j h))) \right) + \mathcal{O}(h^{q+1}).$$

This implies that

$$Y_{ni} - y(x_n + c_i h) = \mathcal{O}(h^\nu) \quad \text{with} \quad \nu = \min(p, q + 1).$$

Because of (1.11b) we get

$$Z_{ni} - z(x_n + c_i h) = G(Y_{ni}) - G(y(x_n + c_i h)) = \mathcal{O}(h^\nu)$$

and Eq. (1.28) becomes

$$\Delta z_{n+1} = \varrho \Delta z_n + \delta_{n+1}, \quad \text{where} \quad \delta_{n+1} = \mathcal{O}(h^\nu). \quad (1.29)$$

Repeated insertion of this formula gives

$$\Delta z_n = \sum_{i=1}^n \varrho^{n-i} \delta_i, \quad (1.30)$$

because $\Delta z_0 = 0$. This proves the statement for $\varrho \neq -1$. For the case $\varrho = -1$ the error Δz_n is a sum of differences $\delta_{j+1} - \delta_j$. Since δ_{n+1} is actually of the form $\delta_{n+1} = d(x_n)h^\nu + \mathcal{O}(h^{\nu+1})$ we have $\delta_{j+1} - \delta_j = \mathcal{O}(h^{\nu+1})$ and the statement also follows in this situation. \square

The order reduction in the z -component (for non stiffly accurate methods) was first studied by Petzold (1986) in a more general context.

Exercises

1. Compute the solutions of the boundary value problems

$$\begin{aligned} \varepsilon y'' + y' + y = 1 \quad \text{respectively} \quad \varepsilon y'' - y' + y = 1 \quad (1.31) \\ y(0) = y(1) = 0, \quad \text{for} \quad \varepsilon > 0. \end{aligned}$$

Observe that the solutions possess, for $\varepsilon \rightarrow 0$, a “boundary layer” on one of the two sides of $[0, 1]$ and that the limit solutions for $\varepsilon = 0$ satisfy

$$y' + y = 1 \quad \text{respectively} \quad -y' + y = 1$$

with one of the two boundary conditions being lost.

VI.2 Multistep Methods

The aim of this section is to study convergence of multistep methods when applied to singular perturbation problems (Runge-Kutta methods will be treated in Sect. VI.3). We are interested in estimates that hold uniformly for $\varepsilon \rightarrow 0$. The results of the previous chapters cannot be applied. Since the Lipschitz constant of the singular perturbation problem (1.5) is of size $\mathcal{O}(\varepsilon^{-1})$, the estimates of Sect. III.4 are useless. Also the one-sided Lipschitz constant is in general $\mathcal{O}(\varepsilon^{-1})$, so that the convergence results of Sect. V.8 can neither be applied. Let us start by considering the reduced problem.

Methods for Index 1 Problems

A multistep method applied to the system $y' = f(y, z)$, $\varepsilon z' = g(y, z)$ gives

$$\sum_{i=0}^k \alpha_i y_{n+i} = h \sum_{i=0}^k \beta_i f(y_{n+i}, z_{n+i}) \quad (2.1a)$$

$$\varepsilon \sum_{i=0}^k \alpha_i z_{n+i} = h \sum_{i=0}^k \beta_i g(y_{n+i}, z_{n+i}). \quad (2.1b)$$

By putting $\varepsilon = 0$ we obtain (ε -embedding method)

$$\sum_{i=0}^k \alpha_i y_{n+i} = h \sum_{i=0}^k \beta_i f(y_{n+i}, z_{n+i}) \quad (2.2a)$$

$$0 = \sum_{i=0}^k \beta_i g(y_{n+i}, z_{n+i}) \quad (2.2b)$$

which allows us to apply a multistep method to the differential-algebraic system (1.6). This approach was first proposed (for the BDF methods) by Gear (1971).

Theorem 2.1. *Suppose that the system (1.6) satisfies (1.7). Consider a multistep method of order p which is stable at the origin and at infinity (0 and ∞ are in the stability region) and suppose that the error of the starting values y_j, z_j for $j = 0, \dots, k-1$ is $\mathcal{O}(h^p)$. Then the global error of (2.2) satisfies*

$$y_n - y(x_n) = \mathcal{O}(h^p), \quad z_n - z(x_n) = \mathcal{O}(h^p)$$

for $x_n - x_0 = nh \leq \text{Const.}$

Proof. Formula (2.2b) is a stable recursion for $\delta_n = g(y_n, z_n)$, because ∞ lies in the stability region of the method. This together with the assumption on the starting values implies that $\delta_n = \mathcal{O}(h^p)$ for all $n \geq 0$. By the Implicit Function Theorem $g(y_n, z_n) = \delta_n$ can be solved for z_n and yields

$$z_n = G(y_n) + \mathcal{O}(h^p) \quad (2.3)$$

with $G(y)$ as in (1.8). Inserting (2.3) into (2.2a) gives the multistep formula for the differential equation (1.8) with an $\mathcal{O}(h^{p+1})$ perturbation. The statement then follows from the convergence proof of Sect. III.4. \square

For the implicit index 1 problem (1.17) the multistep method becomes

$$M \sum_{i=0}^k \alpha_i u_{n+i} = h \sum_{i=0}^k \beta_i \varphi(u_{n+i}) \quad (2.4)$$

and convergence without any order reduction for methods satisfying the hypotheses of Theorem 2.1 follows from the transformation (1.20) and the diagram (1.23).

The *state space from approach* is also possible for multistep methods. We just have to replace (2.2b) by

$$g(y_{n+k}, z_{n+k}) = 0. \quad (2.2c)$$

Method (2.2a,c) is equivalent to the solution of (1.8) by the above multistep method. Hence, we have convergence as for nonstiff ordinary differential equations. The assumption “ $\infty \in S$ ” is no longer necessary and even explicit methods can be applied.

Convergence for Singular Perturbation Problems

The error propagation has been studied by Söderlind & Dahlquist (1981) using G -stability estimates. Convergence results were first obtained by Lötstedt (1985) for BDF methods. The following convergence result by Lubich (1991), based on the smoothness of the exact solution and thus uniform in ε as long as we stay away from transient phases, gives optimal error bounds for arbitrary multistep methods.

The Jacobian of the system (1.5) is of the form

$$\begin{pmatrix} f_y & f_z \\ \varepsilon^{-1}g_y & \varepsilon^{-1}g_z \end{pmatrix}$$

and its dominant eigenvalues are seen to be close to $\varepsilon^{-1}\lambda$ where λ represents the eigenvalues of g_z . For reasons of stability we assume throughout this subsection that the eigenvalues of g_z have negative real part. More precisely, we assume that

$$\text{the eigenvalues } \lambda \text{ of } g_z(y, z) \text{ lie in } |\arg \lambda - \pi| < \alpha \quad (2.5)$$

for (y, z) in a neighbourhood of the considered solution. We then have the following result for method (2.1a,b):

Theorem 2.2 (Lubich 1991). *Suppose that the multistep method is of order p , $A(\alpha)$ -stable and strictly stable at infinity. If the problem (1.5) satisfies (2.5), then the error is bounded for $h \geq \varepsilon$ and $nh \leq \bar{x} - x_0$ by*

$$\begin{aligned} & \|y_n - y(x_n)\| + \|z_n - z(x_n)\| \\ & \leq C \left(\max_{0 \leq j < k} \|y_j - y(x_j)\| + h^p \int_{x_0}^{x_n} \|y^{(p+1)}(x)\| dx \right. \\ & \quad \left. + (h + \varrho^n) \max_{0 \leq j < k} \|z_j - z(x_j)\| + \varepsilon h^p \max_{x_0 \leq x \leq x_n} \|z^{(p+1)}(x)\| \right) \end{aligned}$$

with $0 < \varrho < 1$. This estimate holds for $h \leq h_0$ (h_0 sufficiently small, but independent of ε), and provided that the starting values are in a sufficiently small, h - and ε -independent neighbourhood of the exact solution. The constants C and ϱ are independent of ε and h .

Proof. The proof is divided into several parts: in part (a) we shall derive recursive estimates for the global error, these will be solved in part (b); part (c) proves an inequality which is needed in (a).

(a) First we insert the exact solution of (1.5) into the method (2.1) and so obtain

$$\sum_{i=0}^k \alpha_i y(x_{n+i}) = h \sum_{i=0}^k \beta_i f(y(x_{n+i}), z(x_{n+i})) + d_{n+k} \quad (2.6a)$$

$$\sum_{i=0}^k \alpha_i z(x_{n+i}) = \frac{h}{\varepsilon} \sum_{i=0}^k \beta_i g(y(x_{n+i}), z(x_{n+i})) + e_{n+k}, \quad (2.6b)$$

where the perturbations d_{n+k} , e_{n+k} can be estimated (for $n \geq 0$) as

$$\|d_{n+k}\| \leq C_1 h^p \int_{x_n}^{x_{n+k}} \|y^{(p+1)}(x)\| dx \quad (2.7a)$$

$$\|e_{n+k}\| \leq C_2 h^{p+1} \max_{x_n \leq x \leq x_{n+k}} \|z^{(p+1)}(x)\|. \quad (2.7b)$$

We then denote the global errors by $\Delta y_n = y_n - y(x_n)$, $\Delta z_n = z_n - z(x_n)$ and introduce the differences

$$\Delta f_{n+k} = \sum_{i=0}^k \beta_i \left(f(y_{n+i}, z_{n+i}) - f(y(x_{n+i}), z(x_{n+i})) \right), \quad n \geq 0,$$

$\Delta f_j = 0$ for $j < k$. Subtraction of (2.6a) from (2.1a) yields for $n \geq 0$

$$\sum_{i=0}^k \alpha_i \Delta y_{n+i} = h \Delta f_{n+k} - d_{n+k}. \quad (2.8)$$

Guided by previous experience (see (V.7.41)), we define d_0, \dots, d_{k-1} so that (2.8)

also holds for negative n . Solving for Δy_n gives

$$\Delta y_n = h \sum_{j=0}^n r_{n-j}(0) \Delta f_j - \sum_{j=0}^n r_{n-j}(0) d_j$$

where $r_j(0)$ is defined in (V.7.44). These numbers are the coefficients of $r(\zeta, 0) = \zeta^{-k}/\varrho(\zeta^{-1})$. By zero-stability of the method, the sequence $\{r_j(0)\}$ is bounded, so that a Lipschitz condition for $f(y, z)$ implies the estimate

$$\|\Delta y_n\| \leq h \sum_{j=0}^n (M\|\Delta y_j\| + N\|\Delta z_j\|) + C_3 \sum_{j=0}^n \|d_j\|. \quad (2.9)$$

A more refined estimate is necessary for the z -component. We take the difference of (2.1b) and (2.6b) and then subtract from both sides the quantity

$$\frac{h}{\varepsilon} \sum_{i=0}^k \beta_i J \Delta z_{n+i} \quad \text{where} \quad J = g_z(y_0, z_0). \quad (2.10)$$

This yields

$$\sum_{i=0}^k (\alpha_i I - \beta_i \frac{h}{\varepsilon} J) \Delta z_{n+i} = \frac{h}{\varepsilon} \Delta g_{n+k} - e_{n+k} \quad (2.11)$$

where

$$\Delta g_{n+k} = \sum_{i=0}^k \beta_i \left(g(y_{n+i}, z_{n+i}) - g(y(x_{n+i}), z(x_{n+i})) - J \Delta z_{n+i} \right), \quad (2.12)$$

and $\Delta g_j = 0$ for $j < k$. We again define e_0, \dots, e_{k-1} such that (2.11) holds for negative n , and we then solve (2.11) for Δz_n . This gives

$$\Delta z_n = \frac{h}{\varepsilon} \sum_{j=0}^n r_{n-j} \left(\frac{h}{\varepsilon} J \right) \Delta g_j - \sum_{j=0}^n r_{n-j} \left(\frac{h}{\varepsilon} J \right) e_j \quad (2.13)$$

where the matrices $r_j(\frac{h}{\varepsilon} J)$ are defined by (see Formula (V.7.50))

$$\frac{h}{\varepsilon} \sum_{j \geq 0} r_j \left(\frac{h}{\varepsilon} J \right) \zeta^j = \left(\frac{\varepsilon}{h} \delta(\zeta) I - J \right)^{-1} \frac{\zeta^{-k}}{\sigma(\zeta^{-1})} \quad (2.14)$$

with $\delta(\zeta)$ given in (V.7.45). In part (c) below we shall prove that

$$\frac{h}{\varepsilon} \left\| r_j \left(\frac{h}{\varepsilon} J \right) \right\| \leq C \kappa^j \quad \text{with} \quad 0 < \kappa < 1. \quad (2.15)$$

Inserted into (2.13) we thus get

$$\|\Delta z_n\| \leq \sum_{j=0}^n \kappa^{n-j} (L\|\Delta y_j\| + \ell\|\Delta z_j\|) + C_4 \frac{\varepsilon}{h} \sum_{j=0}^n \kappa^{n-j} \|e_j\|. \quad (2.16)$$

It is important to remark that the Lipschitz constant ℓ can be made arbitrarily small by shrinking the considered interval.

b) In order to solve the inequalities (2.9) and (2.16) we define sequences $\{u_n\}$ and $\{v_n\}$ by

$$\begin{aligned} u_n &= h \sum_{j=0}^n (Mu_j + Nv_j) + C_3 \sum_{j=0}^n \|d_j\|, \\ v_n &= \sum_{j=0}^n \kappa^{n-j} (Lu_j + \ell v_j) + C_4 \frac{\varepsilon}{h} \sum_{j=0}^n \kappa^{n-j} \|e_j\|. \end{aligned} \quad (2.17)$$

An induction argument shows that

$$\|\Delta y_n\| \leq u_n, \quad \|\Delta z_n\| \leq v_n$$

provided $\ell < 1$ and $h \leq h_0$. We then rewrite (2.17) as

$$\begin{aligned} u_n &= u_{n-1} + hMu_n + hNv_n + C_3 \|d_n\|, & u_{-1} &= 0, \\ v_n &= \kappa v_{n-1} + Lu_n + \ell v_n + C_4 \frac{\varepsilon}{h} \|e_n\|, & v_{-1} &= 0. \end{aligned}$$

Solving for u_n, v_n we get (with $\varrho = \kappa/(1-\ell)$)

$$\begin{pmatrix} u_n \\ v_n \end{pmatrix} = A(h) \begin{pmatrix} u_{n-1} \\ v_{n-1} \end{pmatrix} + \begin{pmatrix} \hat{d}_n \\ \hat{e}_n \end{pmatrix}, \quad A(h) = \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & \varrho + \mathcal{O}(h) \end{pmatrix} \quad (2.18)$$

where

$$|\hat{d}_n| \leq C_5 (\|d_n\| + \varepsilon \|e_n\|), \quad |\hat{e}_n| \leq C_6 (\|d_n\| + \frac{\varepsilon}{h} \|e_n\|). \quad (2.19)$$

Inserting (2.18) repeatedly we obtain

$$\begin{pmatrix} u_n \\ v_n \end{pmatrix} = \sum_{j=0}^n A(h)^{n-j} \begin{pmatrix} \hat{d}_j \\ \hat{e}_j \end{pmatrix}. \quad (2.20)$$

If ℓ is small enough so that $\varrho = \kappa/(1-\ell) < 1$ and if $h \leq h_0$, then the eigenvalues of $A(h)$ are distinct and $A(h)$ can be diagonalized as

$$A(h) = T^{-1}(h) \begin{pmatrix} 1 + \mathcal{O}(h) & 0 \\ 0 & \varrho + \mathcal{O}(h) \end{pmatrix} T(h), \quad T(h) = \begin{pmatrix} 1 & \mathcal{O}(h) \\ \mathcal{O}(1) & 1 \end{pmatrix}.$$

Inserted into (2.20) this yields

$$u_n + v_n \leq \text{Const.} \left(\sum_{j=1}^n \hat{d}_j + \sum_{j=1}^n (h + \varrho^{n-j}) \hat{e}_j \right).$$

Since d_0, \dots, d_{k-1} are linear combinations of the values $\Delta y_0, \dots, \Delta y_{k-1}$, and e_0, \dots, e_{k-1} are linear combinations of the Δz_j and $\frac{h}{\varepsilon} \Delta z_j$, the statement of the theorem follows from (2.19) and (2.7). Because of our assumption on ℓ (that $\varrho = \kappa/(1-\ell) < 1$) we have proved the theorem for sufficiently small (but ε -independent) intervals. Compact intervals $[x_0, \bar{x}]$ can be covered by repeated application of the above estimates.

c) It still remains to prove (2.15). More generally, we shall show that

$$\frac{h}{\varepsilon} \left\| r_j \left(\frac{h}{\varepsilon} g_z(y, z) \right) \right\| \leq C \kappa^j \quad \text{with} \quad 0 < \kappa < 1 \quad (2.21)$$

holds uniformly in a compact neighbourhood of the solution. This is necessary, if the above estimates are applied to several subintervals. In order to prove (2.21) we remember that $r_j(\frac{h}{\varepsilon} J)$ is defined by (2.14). If we are able to show that

$$\left\| \left(\frac{\varepsilon}{h} \delta(\zeta) I - g_z(y, z) \right)^{-1} \frac{\zeta^{-k}}{\sigma(\zeta^{-k})} \right\| \leq C \quad \text{for} \quad |\zeta| \leq 1/\kappa \quad (2.22)$$

then the estimate (2.21) follows immediately from Cauchy's integral formula

$$\frac{h}{\varepsilon} r_j \left(\frac{h}{\varepsilon} J \right) = \frac{1}{2\pi i} \int_{|\zeta|=1/\kappa} \left(\frac{\varepsilon}{h} \delta(\zeta) I - J \right)^{-1} \frac{\zeta^{-k}}{\sigma(\zeta^{-1})} \cdot \zeta^{-j-1} d\zeta.$$

By definition of the stability region S of a multistep method, the value $\delta(\zeta)$ lies outside of S whenever $|\zeta| < 1$. Recall that the method is $A(\alpha)$ -stable and strictly stable at infinity, and the differential equation satisfies (2.5). Therefore the set of eigenvalues of $g_z(y, z)$ (with (y, z) varying in a compact neighbourhood of the solution) is well separated from $\{\gamma \delta(\zeta) ; \gamma \leq 1, |\zeta| \leq 1\}$. It is even separated from $\{\gamma \delta(\zeta) ; \gamma \leq 1, |\zeta| \leq 1/\kappa\}$ with some $\kappa < 1$. Together with Exercise 2 of Sect. V.7 this proves (2.22). \square

Exercises

1. (Lubich 1991). Prove that for the BDF-schemes the estimate of Theorem 2.2 (for $n \geq k$) is valid with $(h + \varrho^n)$ replaced by $\varepsilon(1 + \varrho^n/h)$ in the factor multiplying the z -component of the errors in the starting values.

Hint. Give a direct proof for $n \in \{k, \dots, 2k-1\}$; then apply Theorem 2.2 to shifted starting values.

VI.3 Epsilon Expansions for Exact and RK Solutions

In the preceding section we have proved convergence of multistep methods for singular perturbation problems. The same techniques do not yield optimal estimates for Runge-Kutta methods. We therefore investigate more thoroughly the structure of the solutions of singular perturbation problems. A first systematic study of the qualitative aspects of such problems is due to Tikhonov (1952). Asymptotic expansions were then analyzed by Vasil'eva (1963). Classical books on this subject are Wasow (1965), O'Malley (1974), and Tikhonov, Vasil'eva & Sveshnikov (1985).

Expansion of the Smooth Solution

Tikhonov's theorem is only the first step . . . The actual approximate solution of such problems in series form is still a difficult question. It has been analyzed in a series of papers by Vasil'eva . . .
(W. Wasow 1965)

We consider the singular perturbation problem

$$\begin{aligned} y' &= f(y, z) \\ \varepsilon z' &= g(y, z), \quad 0 < \varepsilon \ll 1 \end{aligned} \tag{3.1}$$

where f and g are sufficiently differentiable. The functions f, g and the initial values $y(0), z(0)$ may depend smoothly on ε . For simplicity of notation we suppress this dependence. The corresponding equation for $\varepsilon = 0$,

$$\begin{aligned} y' &= f(y, z) \\ 0 &= g(y, z), \end{aligned} \tag{3.2}$$

is the *reduced problem*. In order to guarantee the solvability of (3.2), we assume that $g_z(y, z)$ is invertible (in a neighbourhood of the solution of (3.2)).

We are mainly interested in smooth solutions of (3.1), which are of the form

$$\begin{aligned} y(x) &= y_0(x) + \varepsilon y_1(x) + \varepsilon^2 y_2(x) + \dots \\ z(x) &= z_0(x) + \varepsilon z_1(x) + \varepsilon^2 z_2(x) + \dots \end{aligned} \tag{3.3}$$

Inserting (3.3) into (3.1) and collecting equal powers of ε yields

$$\varepsilon^0 : \quad \left. \begin{aligned} y'_0 &= f(y_0, z_0) \\ 0 &= g(y_0, z_0) \end{aligned} \right\} \tag{3.4a}$$

$$\varepsilon^1 : \quad \left. \begin{aligned} y_1' &= f_y(y_0, z_0)y_1 + f_z(y_0, z_0)z_1 \\ z_0' &= g_y(y_0, z_0)y_1 + g_z(y_0, z_0)z_1 \end{aligned} \right\} \quad (3.4b)$$

$$\varepsilon^\nu : \quad \left. \begin{aligned} y_\nu' &= f_y(y_0, z_0)y_\nu + f_z(y_0, z_0)z_\nu + \varphi_\nu(y_0, z_0, \dots, y_{\nu-1}, z_{\nu-1}) \\ z_{\nu-1}' &= g_y(y_0, z_0)y_\nu + g_z(y_0, z_0)z_\nu + \psi_\nu(y_0, z_0, \dots, y_{\nu-1}, z_{\nu-1}) \end{aligned} \right\} \quad (3.4c)$$

As expected, we see from (3.4a) that $y_0(x)$, $z_0(x)$ is a solution of the reduced system. Since g_z is invertible, the second equation of (3.4b) can be solved for z_1 . By inserting z_1 into the upper relation of (3.4b) we obtain a linear differential equation for $y_1(x)$. Hence, $y_1(x)$ and $z_1(x)$ are determined. Similarly, we get $y_2(x)$, $z_2(x)$ from (3.4c), etc.

This construction of the coefficients of (3.3) shows that we can choose the initial values $y_j(0)$ arbitrarily, but that there is no freedom in the choice of $z_j(0)$. Consequently, not every solution of (3.1) can be written in the form (3.3).

Expansions with Boundary Layer Terms

To construct a uniform asymptotic expansion we must combine the Maclaurin expansion with another expansion of special form. The terms in this expansion are exponential functions that are appreciable inside the boundary layer, but negligibly small outside it. (A.B. Vasil'eva 1963)

Example 3.1. We consider the problem (IV.1.1), written in the form

$$\varepsilon z' = -z + \cos x. \quad (3.5)$$

Its analytic solution

$$\begin{aligned} z(x) &= (1 + \varepsilon^2)^{-1} (\cos x + \varepsilon \sin x) + C e^{-x/\varepsilon} \\ &= \cos x + \varepsilon \sin x - \varepsilon^2 \cos x - \varepsilon^3 \sin x + \dots + C e^{-x/\varepsilon} \end{aligned}$$

is a superposition of a smooth solution of the form (3.3) and of a rapidly decaying function. This additional term (transient phase, boundary layer) compensates the missing freedom in the choice of the initial values $z_j(0)$.

Motivated by this example, we seek solutions of the general problem (3.1) which are of the form

$$\begin{aligned} y(x) &= \sum_{j \geq 0} \varepsilon^j y_j(x) + \varepsilon \sum_{j \geq 0} \varepsilon^j \eta_j(x/\varepsilon) \\ z(x) &= \sum_{j \geq 0} \varepsilon^j z_j(x) + \sum_{j \geq 0} \varepsilon^j \zeta_j(x/\varepsilon), \end{aligned} \quad (3.6)$$

where $y_j(x)$, $z_j(x)$ are determined by (3.4) and the ε -independent functions η_j , ζ_j are assumed to satisfy

$$\|\eta_j(\xi)\| \leq e^{-\kappa \xi}, \quad \|\zeta_j(\xi)\| \leq e^{-\kappa \xi} \quad (3.7)$$

with some $\kappa > 0$. Inserting (3.6) into (3.1) and using (3.4) we obtain formally

$$\begin{aligned} \sum_{j \geq 0} \varepsilon^j \eta_j' \left(\frac{x}{\varepsilon} \right) &= f \left(\sum_{j \geq 0} \varepsilon^j y_j(x) + \varepsilon \sum_{j \geq 0} \varepsilon^j \eta_j \left(\frac{x}{\varepsilon} \right), \sum_{j \geq 0} \varepsilon^j z_j(x) + \sum_{j \geq 0} \varepsilon^j \zeta_j \left(\frac{x}{\varepsilon} \right) \right) \\ &\quad - f \left(\sum_{j \geq 0} \varepsilon^j y_j(x), \sum_{j \geq 0} \varepsilon^j z_j(x) \right) \end{aligned} \quad (3.8a)$$

$$\begin{aligned} \sum_{j \geq 0} \varepsilon^j \zeta_j' \left(\frac{x}{\varepsilon} \right) &= g \left(\sum_{j \geq 0} \varepsilon^j y_j(x) + \varepsilon \sum_{j \geq 0} \varepsilon^j \eta_j \left(\frac{x}{\varepsilon} \right), \sum_{j \geq 0} \varepsilon^j z_j(x) + \sum_{j \geq 0} \varepsilon^j \zeta_j \left(\frac{x}{\varepsilon} \right) \right) \\ &\quad - g \left(\sum_{j \geq 0} \varepsilon^j y_j(x), \sum_{j \geq 0} \varepsilon^j z_j(x) \right). \end{aligned} \quad (3.8b)$$

We then replace x by the stretched variable

$$\xi = x/\varepsilon \quad (3.9)$$

and compare like powers of ε in (3.8). This gives for ε^0

$$\eta_0'(\xi) = f(y_0(0), z_0(0) + \zeta_0(\xi)) - f(y_0(0), z_0(0)) \quad (3.10a)$$

$$\zeta_0'(\xi) = g(y_0(0), z_0(0) + \zeta_0(\xi)) - g(y_0(0), z_0(0)). \quad (3.10b)$$

At this point it is necessary to introduce some stability assumption for (3.1) in order to obtain (3.7). We shall require that the logarithmic norm of g_z satisfy

$$\mu(g_z(y, z)) \leq -1 \quad (3.11)$$

in an ε -independent neighbourhood of the solution of (3.2) (any negative bound other than -1 can be normalized by re-scaling ε). By Theorem I.10.6 Eqs. (3.10b) and (3.11) imply

$$\|\zeta_0(\xi)\| \leq \|\zeta_0(0)\| e^{-\xi}.$$

Since $f(y, z)$ satisfies locally a Lipschitz condition, the right-hand side of (3.10a), denoted by $\varphi(\xi)$, is bounded by $\|\varphi(\xi)\| \leq L \|\zeta_0(0)\| e^{-\xi}$. Consequently, there is only one solution of (3.10a) which satisfies (3.7), namely

$$\eta_0(\xi) = \int_0^\xi \varphi(s) ds - \int_0^\infty \varphi(s) ds. \quad (3.12)$$

A comparison of the powers of ε^1 in (3.8) yields

$$\begin{aligned} \eta_1'(\xi) &= f_y(y_0(0), z_0(0) + \zeta_0(\xi)) (y_1(0) + \xi y_0'(0) + \eta_0(\xi)) \\ &\quad + f_z(y_0(0), z_0(0) + \zeta_0(\xi)) (z_1(0) + \xi z_0'(0) + \zeta_1(\xi)) \\ &\quad - f_y(y_0(0), z_0(0)) (y_1(0) + \xi y_0'(0)) \\ &\quad - f_z(y_0(0), z_0(0)) (z_1(0) + \xi z_0'(0)) \\ &\quad \zeta_1'(\xi) = g_y(y_0(0), z_0(0) + \zeta_0(\xi)) (y_1(0) + \xi y_0'(0) + \eta_0(\xi)) \\ &\quad + g_z(y_0(0), z_0(0) + \zeta_0(\xi)) (z_1(0) + \xi z_0'(0) + \zeta_1(\xi)) \end{aligned} \quad (3.13a)$$

$$\begin{aligned}
& -g_y(y_0(0), z_0(0))(y_1(0) + \xi y'_0(0)) \\
& -g_z(y_0(0), z_0(0))(z_1(0) + \xi z'_0(0)).
\end{aligned} \tag{3.13b}$$

Eq. (3.13b) is a linear differential equation for $\zeta_1(\xi)$. Its defect, for ζ_1 replaced by 0, is bounded by $Ce^{-\xi}$. Therefore, an application of Theorem I.10.6 yields

$$\|\zeta_1(\xi)\| \leq e^{-\xi}(\|\zeta_1(0)\| + C\xi),$$

which implies (3.7) for any $\kappa < 1$. The right-hand side of (3.13a) is then bounded by $C_1 e^{-\kappa\xi}$. As in (3.12) we obtain a unique solution to (3.13a), which satisfies (3.7). This procedure can be continued to construct all further $\eta_j(\xi)$, $\zeta_j(\xi)$. At each step, the value of κ in (3.7) may become smaller. This is no serious difficulty, because we are only interested in a finite part of the series (3.6).

We point out that for the construction of $\eta_j(\xi)$, $\zeta_j(\xi)$ we can choose $\zeta_j(0)$ arbitrarily, but that there is no freedom in the choice of $\eta_j(0)$.

As a consequence, for an arbitrary initial value for (3.1) with expansion

$$\begin{aligned}
y(0) &= y_0^0 + \varepsilon y_1^0 + \varepsilon^2 y_2^0 + \dots \\
z(0) &= z_0^0 + \varepsilon z_1^0 + \varepsilon^2 z_2^0 + \dots,
\end{aligned} \tag{3.14}$$

the coefficients of the series (3.6) can be constructed as follows: put $x = 0$ in (3.6) to obtain the necessary relations

$$y_0(0) = y_0^0, \quad y_j(0) + \eta_{j-1}(0) = y_j^0, \quad z_j(0) + \zeta_j(0) = z_j^0. \tag{3.15}$$

This initial value $y_0(0) = y_0^0$ determines $z_0(0)$ by (3.4a), $\zeta_0(0)$ is then given by (3.15), $\eta_0(0)$ by (3.12), $y_1(0)$ by (3.15), $z_1(0)$ by (3.4b), $\zeta_1(0)$ by (3.15), $\eta_1(0)$ by (3.13a) and (3.7), $y_2(0)$ by (3.15), etc.

Estimation of the Remainder

The following result gives a rigorous estimate of the remainder in (3.6), when only a truncated series is considered.

Theorem 3.2. *Consider the initial value problem (3.1), (3.14), and suppose that (3.11) holds in an ε -independent neighbourhood of the solution $y_0(x)$, $z_0(x)$ ($0 \leq x \leq \bar{x}$) of the reduced problem ($y_0(0) = y_0^0$). If (y_0^0, z_0^0) lies in this neighbourhood, then the problem (3.1), (3.14) has a unique solution for ε sufficiently small and for $0 \leq x \leq \bar{x}$, which is of the form*

$$\begin{aligned}
y(x) &= \sum_{j=0}^N \varepsilon^j y_j(x) + \varepsilon \sum_{j=0}^{N-1} \varepsilon^j \eta_j(x/\varepsilon) + \mathcal{O}(\varepsilon^{N+1}) \\
z(x) &= \sum_{j=0}^N \varepsilon^j z_j(x) + \sum_{j=0}^N \varepsilon^j \zeta_j(x/\varepsilon) + \mathcal{O}(\varepsilon^{N+1}).
\end{aligned} \tag{3.16}$$

The coefficients $y_j(x)$, $z_j(x)$, $\eta_j(\xi)$, $\zeta_j(\xi)$ are given by (3.4), (3.10), (3.13), and satisfy (3.7).

Proof. We denote the truncated series by

$$\begin{aligned}\widehat{y}(x) &= \sum_{j=0}^N \varepsilon^j y_j(x) + \varepsilon \sum_{j=0}^N \varepsilon^j \eta_j(x/\varepsilon) \\ \widehat{z}(x) &= \sum_{j=0}^N \varepsilon^j z_j(x) + \sum_{j=0}^N \varepsilon^j \zeta_j(x/\varepsilon).\end{aligned}\tag{3.17}$$

By our construction of $y_j(x)$, $z_j(x)$, $\eta_j(\xi)$, $\zeta_j(\xi)$ we have

$$\begin{aligned}\widehat{y}'(x) &= f(\widehat{y}(x), \widehat{z}(x)) + \mathcal{O}(\varepsilon^{N+1}) \\ \varepsilon \widehat{z}'(x) &= g(\widehat{y}(x), \widehat{z}(x)) + \mathcal{O}(\varepsilon^{N+1}).\end{aligned}\tag{3.18}$$

Subtracting (3.1) from (3.18) and exploiting Lipschitz conditions for f and g we obtain

$$\begin{aligned}D_+ \|\widehat{y}(x) - y(x)\| &\leq L_1 \|\widehat{y}(x) - y(x)\| + L_2 \|\widehat{z}(x) - z(x)\| + C_1 \varepsilon^{N+1} \\ \varepsilon D_+ \|\widehat{z}(x) - z(x)\| &\leq L_3 \|\widehat{y}(x) - y(x)\| - \|\widehat{z}(x) - z(x)\| + C_2 \varepsilon^{N+1}.\end{aligned}\tag{3.19}$$

Here, D_+ denotes the Dini derivate introduced in Section I.10. We have used $D_+ \|w(x)\| \leq \|w'(x)\|$ (see Eq. (I.10.4)) and, for the second inequality of (3.19), Formula (I.10.17) together with (3.11).

In order to solve inequality (3.19) we replace \leq by $=$ and so obtain

$$\begin{aligned}u' &= L_1 u + L_2 v + C_1 \varepsilon^{N+1}, & u_0 &= \|\widehat{y}(0) - y(0)\| = \mathcal{O}(\varepsilon^{N+1}) \\ \varepsilon v' &= L_3 u - v + C_2 \varepsilon^{N+1}, & v_0 &= \|\widehat{z}(0) - z(0)\| = \mathcal{O}(\varepsilon^{N+1}).\end{aligned}\tag{3.20}$$

This system is quasimonotone, it thus follows from Exercise 7 (Sect. I.10) that

$$\|\widehat{y}(x) - y(x)\| \leq u(x), \quad \|\widehat{z}(x) - z(x)\| \leq v(x).\tag{3.21}$$

Transforming (3.20) to diagonal form one easily finds its analytic solution and verifies that $u(x) = \mathcal{O}(\varepsilon^{N+1})$, $v(x) = \mathcal{O}(\varepsilon^{N+1})$ on compact intervals. Inserted into (3.21) this proves the statement. \square

Expansion of the Runge-Kutta Solution

After having understood the structure of the analytic solution of (3.1), we turn our attention to its numerical counterpart. We consider the Runge-Kutta method

$$\begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} y_n \\ z_n \end{pmatrix} + h \sum_{i=1}^s b_i \begin{pmatrix} k_{ni} \\ \ell_{ni} \end{pmatrix}\tag{3.22}$$

where

$$\begin{pmatrix} k_{ni} \\ \varepsilon \ell_{ni} \end{pmatrix} = \begin{pmatrix} f(Y_{ni}, Z_{ni}) \\ g(Y_{ni}, Z_{ni}) \end{pmatrix}\tag{3.23}$$

and the internal stages are given by

$$\begin{pmatrix} Y_{ni} \\ Z_{ni} \end{pmatrix} = \begin{pmatrix} y_n \\ z_n \end{pmatrix} + h \sum_{j=1}^s a_{ij} \begin{pmatrix} k_{nj} \\ \ell_{nj} \end{pmatrix}. \quad (3.24)$$

For arbitrary initial values, the solution possesses a transient phase (as described by Theorem 3.2), and the numerical method has anyway to take small step sizes of magnitude $\mathcal{O}(\varepsilon)$. We shall therefore focus on the situation where the transient phase is over and the method has reached the smooth solution within the given tolerance. We thus suppose that the initial values lie on the smooth solution (i.e., that an expansion of the form (3.3) holds) and that the step size h is large compared to ε . Our first goal is an ε -expansion of the numerical solution. To this end, we formally expand all occurring quantities into powers of ε with ε -independent coefficients (see Hairer, Lubich & Roche 1988)

$$y_n = y_n^0 + \varepsilon y_n^1 + \varepsilon^2 y_n^2 + \dots \quad (3.25a)$$

$$Y_{ni} = Y_{ni}^0 + \varepsilon Y_{ni}^1 + \varepsilon^2 Y_{ni}^2 + \dots \quad (3.25b)$$

$$k_{ni} = k_{ni}^0 + \varepsilon k_{ni}^1 + \varepsilon^2 k_{ni}^2 + \dots \quad (3.25c)$$

and similarly for z_n, Z_{ni}, ℓ_{ni} . Because of the linearity of the relations (3.22) and (3.24) we have

$$\begin{pmatrix} y_{n+1}^\nu \\ z_{n+1}^\nu \end{pmatrix} = \begin{pmatrix} y_n^\nu \\ z_n^\nu \end{pmatrix} + h \sum_{i=1}^s b_i \begin{pmatrix} k_{ni}^\nu \\ \ell_{ni}^\nu \end{pmatrix} \quad (3.26)$$

and

$$\begin{pmatrix} Y_{ni}^\nu \\ Z_{ni}^\nu \end{pmatrix} = \begin{pmatrix} y_n^\nu \\ z_n^\nu \end{pmatrix} + h \sum_{j=1}^s a_{ij} \begin{pmatrix} k_{nj}^\nu \\ \ell_{nj}^\nu \end{pmatrix}. \quad (3.27)$$

Inserting (3.25b, c) into (3.23) and comparing equal powers of ε we obtain

$$\varepsilon^0 : \left. \begin{aligned} k_{ni}^0 &= f(Y_{ni}^0, Z_{ni}^0) \\ 0 &= g(Y_{ni}^0, Z_{ni}^0) \end{aligned} \right\} \quad (3.28a)$$

$$\varepsilon^1 : \left. \begin{aligned} k_{ni}^1 &= f_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^1 + f_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^1 \\ \ell_{ni}^0 &= g_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^1 + g_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^1 \end{aligned} \right\} \quad (3.28b)$$

$$\dots$$

$$\varepsilon^\nu : \left. \begin{aligned} k_{ni}^\nu &= f_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^\nu + f_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^\nu + \varphi_\nu(Y_{ni}^0, Z_{ni}^0, \dots, Y_{ni}^{\nu-1}, Z_{ni}^{\nu-1}) \\ \ell_{ni}^{\nu-1} &= g_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^\nu + g_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^\nu + \psi_\nu(Y_{ni}^0, Z_{ni}^0, \dots, Y_{ni}^{\nu-1}, Z_{ni}^{\nu-1}) \end{aligned} \right\} \quad (3.28c)$$

Since (3.23) has the same form as the differential equation (3.1), it is obvious that the formulas of (3.28) are exactly the same as those of (3.4). An interesting interpretation of this fact is the following: the coefficients $y_n^0, z_n^0, y_n^1, z_n^1, \dots$ represent the numerical solution of the Runge-Kutta method applied to the differential-algebraic system (3.4) (ε -embedding method of Sect. VI.1). This can be expressed

by the commutativity of the following diagram:

$$\begin{array}{ccc}
 \text{Problem (3.1)} & \xrightarrow{(3.3)} & \text{DAE (3.4)} \\
 \text{RK} \downarrow \text{method} & & \text{RK} \downarrow \text{method} \\
 \{y_n, z_n\} & \xrightarrow{(3.25)} & \{y_n^0, z_n^0, y_n^1, z_n^1, \dots\}
 \end{array}$$

Subtracting (3.25a) from (3.3) we get formally

$$\begin{aligned}
 y_n - y(x_n) &= \sum_{\nu \geq 0} \varepsilon^\nu (y_n^\nu - y_\nu(x_n)) \\
 z_n - z(x_n) &= \sum_{\nu \geq 0} \varepsilon^\nu (z_n^\nu - z_\nu(x_n)).
 \end{aligned} \tag{3.29}$$

In order to study this error we first investigate the differences $y_n^\nu - y_\nu(x_n)$, $z_n^\nu - z_\nu(x_n)$ (next subsection). A rigorous estimate of the remainder in (3.29) will then follow. The presentation follows that of Hairer, Lubich & Roche (1988).

Convergence of RK-Methods for Differential-Algebraic Systems

The first differences $y_n^0 - y_0(x_n)$, $z_n^0 - z_0(x_n)$ in the expansions of (3.29) are just the global errors of the Runge-Kutta method applied to the reduced system (3.4a). By assumption (3.11) this system is of index 1. Therefore, the following result is an immediate consequence of Theorem 1.1.

Theorem 3.3. *Consider a Runge-Kutta method of (classical) order p , with invertible coefficient matrix (a_{ij}) . Suppose that Problem (3.4a) satisfies (3.11) and that the initial values are consistent.*

a) If the method is stiffly accurate (i.e., $a_{si} = b_i$ for $i = 1, \dots, s$) then the global error satisfies

$$y_n^0 - y_0(x_n) = \mathcal{O}(h^p), \quad z_n^0 - z_0(x_n) = \mathcal{O}(h^p). \tag{3.30}$$

b) If the stability function satisfies $|R(\infty)| < 1$, and the stage order is q ($q < p$), then

$$y_n^0 - y_0(x_n) = \mathcal{O}(h^p), \quad z_n^0 - z_0(x_n) = \mathcal{O}(h^{q+1}). \tag{3.31}$$

In both cases the estimates hold uniformly for $nh \leq \text{Const.}$ \square

Estimating the second differences $y_n^1 - y_1(x_n)$, $z_n^1 - z_1(x_n)$ is not as simple, because the enlarged system (3.4a,b) with differential variables y_0, z_0, y_1 and algebraic variable z_1 , is no longer of index 1. It is actually of index 2, as will become clear in Sect. VII.1 below (Exercise 5). In principle it is possible to consult the results of Sect. VII.4 (Theorems VII.4.5 and VII.4.6). For the special system (3.4a,b),

however, a simpler proof is possible. It also extends more easily to the higher-index problems (3.4a-c).

Theorem 3.4 (Hairer, Lubich & Roche 1988). *Consider a Runge-Kutta method of order p , stage order q ($q < p$), such that (a_{ij}) is invertible and the stability function satisfies $|R(\infty)| < 1$. If (3.11) holds and if the initial values of the differential-algebraic system (3.4a-c) are consistent, then the global error of method (3.26)–(3.28) satisfies for $1 \leq \nu \leq q+1$*

$$y_n^\nu - y_\nu(x_n) = \mathcal{O}(h^{q+2-\nu}), \quad z_n^\nu - z_\nu(x_n) = \mathcal{O}(h^{q+1-\nu}).$$

Proof. We denote the differences to the exact solution values by

$$\begin{aligned} \Delta y_n^\nu &= y_n^\nu - y_\nu(x_n), & \Delta z_n^\nu &= z_n^\nu - z_\nu(x_n), \\ \Delta Y_{ni}^\nu &= Y_{ni}^\nu - y_\nu(x_n + c_i h), & \Delta Z_{ni}^\nu &= Z_{ni}^\nu - z_\nu(x_n + c_i h), \\ \Delta k_{ni}^\nu &= k_{ni}^\nu - y'_\nu(x_n + c_i h), & \Delta \ell_{ni}^\nu &= \ell_{ni}^\nu - z'_\nu(x_n + c_i h). \end{aligned} \quad (3.32)$$

Since the quadrature formula with nodes c_i and weights b_i is of order p , we have from (3.26)

$$\begin{pmatrix} \Delta y_{n+1}^\nu \\ \Delta z_{n+1}^\nu \end{pmatrix} = \begin{pmatrix} \Delta y_n^\nu \\ \Delta z_n^\nu \end{pmatrix} + h \sum_{i=1}^s b_i \begin{pmatrix} \Delta k_{ni}^\nu \\ \Delta \ell_{ni}^\nu \end{pmatrix} + \mathcal{O}(h^{p+1}). \quad (3.33)$$

Similarly, the definition of the stage order implies

$$\begin{pmatrix} \Delta Y_{ni}^\nu \\ \Delta Z_{ni}^\nu \end{pmatrix} = \begin{pmatrix} \Delta y_n^\nu \\ \Delta z_n^\nu \end{pmatrix} + h \sum_{j=1}^s a_{ij} \begin{pmatrix} \Delta k_{nj}^\nu \\ \Delta \ell_{nj}^\nu \end{pmatrix} + \mathcal{O}(h^{q+1}). \quad (3.34)$$

It follows from Theorem 3.3 (see also the proof of Theorem 1.1) that

$$\begin{aligned} \Delta y_n^0 &= \mathcal{O}(h^p), & \Delta Y_{ni}^0 &= \mathcal{O}(h^{q+1}), & \Delta k_{ni}^0 &= \mathcal{O}(h^{q+1}) \\ \Delta z_n^0 &= \mathcal{O}(h^{q+1}), & \Delta Z_{ni}^0 &= \mathcal{O}(h^{q+1}), & \Delta \ell_{ni}^0 &= \mathcal{O}(h^q). \end{aligned} \quad (3.35)$$

a) We first consider the case $\nu = 1$. Replacing in (3.28b) Y_{ni}^0, Z_{ni}^0 by $y_0(x_n + c_i h) + \Delta Y_{ni}^0, z_0(x_n + c_i h) + \Delta Z_{ni}^0$ and subtracting Equation (3.4b) at the position $x = x_n + c_i h$, we obtain with the help of (3.35)

$$\begin{aligned} \Delta k_{ni}^1 &= f_y(x_n + c_i h) \Delta Y_{ni}^1 + f_z(x_n + c_i h) \Delta Z_{ni}^1 \\ &\quad + \mathcal{O}(h^{q+1} + h^{q+1} \|\Delta Y_{ni}^1\| + h^{q+1} \|\Delta Z_{ni}^1\|) \\ \Delta \ell_{ni}^0 &= g_y(x_n + c_i h) \Delta Y_{ni}^1 + g_z(x_n + c_i h) \Delta Z_{ni}^1 \\ &\quad + \mathcal{O}(h^{q+1} + h^{q+1} \|\Delta Y_{ni}^1\| + h^{q+1} \|\Delta Z_{ni}^1\|). \end{aligned} \quad (3.36)$$

Here we have used the abbreviations $f_y(x) = f_y(y_0(x), z_0(x))$, etc. Computing ΔZ_{ni}^1 from the second relation of (3.36) and inserting it into the first one yields

$$\begin{aligned} \Delta k_{ni}^1 &- (f_z g_z^{-1})(x_n + c_i h) \Delta \ell_{ni}^0 \\ &= (f_y - f_z g_z^{-1} g_y)(x_n + c_i h) \Delta Y_{ni}^1 + \mathcal{O}(h^{q+1} + h^{q+1} \|\Delta Y_{ni}^1\|). \end{aligned}$$

Using (3.34) we can eliminate ΔY_{ni}^1 and obtain (with (3.35))

$$\Delta k_{ni}^1 - (f_z g_z^{-1})(x_n + c_i h) \Delta \ell_{ni}^0 = \mathcal{O}(\|\Delta y_n^1\|) + \mathcal{O}(h^{q+1}). \quad (3.37)$$

Since $\Delta \ell_{ni}^0$ is of size $\mathcal{O}(h^q)$, we only have $\Delta k_{ni}^1 = \mathcal{O}(\|\Delta y_n^1\|) + \mathcal{O}(h^q)$, and a direct estimation of Δy_n^1 in (3.33) would lead to $\Delta y_n^1 = \mathcal{O}(h^q)$, which is not optimal. We therefore introduce the new variable

$$\Delta u_n^1 = \Delta y_n^1 - (f_z g_z^{-1})(x_n) \Delta z_n^0. \quad (3.38)$$

From (3.33) we get

$$\begin{aligned} \Delta u_{n+1}^1 &= \Delta u_n^1 + h \sum_{i=1}^s b_i (\Delta k_{ni}^1 - (f_z g_z^{-1})(x_n) \Delta \ell_{ni}^0) \\ &\quad - ((f_z g_z^{-1})(x_n + h) - (f_z g_z^{-1})(x_n)) \Delta z_{n+1}^0 + \mathcal{O}(h^{p+1}). \end{aligned} \quad (3.39)$$

The estimates (3.35), (3.37) and the fact that $\Delta y_n^1 = \Delta u_n^1 + \mathcal{O}(h^{q+1})$ imply that

$$\|\Delta u_{n+1}^1\| \leq (1 + Ch) \|\Delta u_n^1\| + \mathcal{O}(h^{q+2}). \quad (3.40)$$

Standard techniques now show that $\Delta u_n^1 = \mathcal{O}(h^{q+1})$ for $nh \leq \text{Const}$ (observe that the initial values are assumed to be consistent, i.e., $\Delta u_0^1 = 0$), so that by (3.38) and (3.35) also $\Delta y_n^1 = \mathcal{O}(h^{q+1})$. This implies $\Delta k_{ni}^1 = \mathcal{O}(h^q)$ by (3.37) and $\Delta Y_{ni}^1 = \mathcal{O}(h^{q+1})$ by (3.34). The second relation of (3.36) then proves that $\Delta Z_{ni}^1 = \mathcal{O}(h^q)$. In order to estimate Δz_n^1 , we compute $\Delta \ell_{ni}^1$ from (3.34) and insert it into (3.33). Using $\Delta Z_{ni}^1 = \mathcal{O}(h^q)$ this gives

$$\Delta z_{n+1}^1 = (1 - b^T A^{-1} \mathbb{1}) \Delta z_n^1 + \mathcal{O}(h^q), \quad (3.41)$$

and it follows from $|1 - b^T A^{-1} \mathbb{1}| = |R(\infty)| < 1$ that $\Delta z_n^1 = \mathcal{O}(h^q)$. We have thus proved the case $\nu = 1$.

b) The proof for general ν is by induction. We shall show that

$$\begin{aligned} \Delta y_n^\nu &= \mathcal{O}(h^{q+2-\nu}), & \Delta Y_{ni}^\nu &= \mathcal{O}(h^{q+2-\nu}) \\ \Delta z_n^\nu &= \mathcal{O}(h^{q+1-\nu}), & \Delta Z_{ni}^\nu &= \mathcal{O}(h^{q+1-\nu}) \end{aligned} \quad (3.42)$$

holds for $\nu = 1, \dots, q+1$. The main difference to the case $\nu = 1$ consists in the additional inhomogeneities φ_ν and ψ_ν in (3.4c). Using their Lipschitz continuity one obtains an additional term of size $\mathcal{O}(h^{q+2-\nu})$ in (3.36). Otherwise the proof is identical to that for $\nu = 1$. \square

We next study the existence and local uniqueness of the solution of the Runge-Kutta method (3.22)–(3.24). Further, we investigate the influence of perturbations in (3.24) to the numerical solution. This will be important for the estimation of the remainder in the expansion (3.29).

Existence and Uniqueness of the Runge-Kutta Solution

For h small compared to ε , the existence of a unique numerical solution of (3.23), (3.24) follows from standard fixed point iteration (e.g., Theorem II.7.2). For the (more interesting) case where the step size h is large compared to ε , we suppose that (y_n, z_n) are known, denote it by (η, ζ) , and prove the existence of (y_{n+1}, z_{n+1}) as follows:

Theorem 3.5 (Hairer, Lubich & Roche 1988). *Assume that $g(\eta, \zeta) = \mathcal{O}(h)$, $\mu(g_z(\eta, \zeta)) \leq -1$ and that the eigenvalues of the Runge-Kutta matrix (a_{ij}) have positive real part. Then, the nonlinear system*

$$\begin{pmatrix} Y_i - \eta \\ \varepsilon(Z_i - \zeta) \end{pmatrix} = h \sum_{j=1}^s a_{ij} \begin{pmatrix} f(Y_j, Z_j) \\ g(Y_j, Z_j) \end{pmatrix} \quad (3.43)$$

possesses a locally unique solution for $h \leq h_0$, where h_0 is sufficiently small but independent of ε . This solution satisfies

$$Y_i - \eta = \mathcal{O}(h), \quad Z_i - \zeta = \mathcal{O}(h). \quad (3.44)$$

Proof. We apply Newton's method to the nonlinear system (3.43), whose second equation is divided by h . The existence and uniqueness statement can then be deduced from the theorem of Newton-Kantorovich (Kantorovich & Akilov 1959, Ortega & Rheinboldt 1970) as follows: for the starting values $Y_i^{(0)} = \eta$, $Z_i^{(0)} = \zeta$ the Jacobian of the system is of the form

$$\begin{pmatrix} I + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & (\varepsilon/h)I - A \otimes g_z(\eta, \zeta) \end{pmatrix}. \quad (3.45)$$

Since $\mu(g_z(\eta, \zeta)) \leq -1$ it follows from the matrix-valued theorem of von Neumann (Theorem V.7.8) that

$$\|(\kappa I - A \otimes g_z(\eta, \zeta))^{-1}\| \leq \max_{\operatorname{Re} \mu \leq -1} \|(\kappa I - \mu A)^{-1}\|. \quad (3.46)$$

The right-hand side of (3.46) is bounded by a constant independent of $\kappa \geq 0$, because the eigenvalues of A are assumed to have positive real part. Consequently, also the inverse of (3.45) is uniformly bounded for $\varepsilon > 0$ and $h \leq h_0$. This together with $g(\eta, \zeta) = \mathcal{O}(h)$ implies that the first increment (of Newton's method) is of size $\mathcal{O}(h)$. Hence, for sufficiently small h , the Newton-Kantorovich assumptions are fulfilled. \square

Influence of Perturbations

For the perturbed Runge-Kutta method

$$\begin{pmatrix} \widehat{Y}_i - \widehat{\eta} \\ \varepsilon(\widehat{Z}_i - \widehat{\zeta}) \end{pmatrix} = h \sum_{j=1}^s a_{ij} \begin{pmatrix} f(\widehat{Y}_j, \widehat{Z}_j) \\ g(\widehat{Y}_j, \widehat{Z}_j) \end{pmatrix} + h \begin{pmatrix} \delta_i \\ \theta_i \end{pmatrix} \quad (3.47)$$

we have the following result.

Theorem 3.6 (Hairer, Lubich & Roche 1988). *Let Y_i, Z_i be given by (3.43) and consider perturbed values $\widehat{Y}_i, \widehat{Z}_i$ satisfying (3.47). In addition to the assumptions of Theorem 3.5 suppose that $\widehat{\eta} - \eta = \mathcal{O}(h)$, $\widehat{\zeta} - \zeta = \mathcal{O}(h)$, $\delta_i = \mathcal{O}(1)$, and $\theta_i = \mathcal{O}(h)$. Then we have for $h \leq h_0$*

$$\begin{aligned} \|\widehat{Y}_i - Y_i\| &\leq C(\|\widehat{\eta} - \eta\| + \varepsilon\|\widehat{\zeta} - \zeta\|) + hC(\|\delta\| + \|\theta\|) \\ \|\widehat{Z}_i - Z_i\| &\leq C(\|\widehat{\eta} - \eta\| + \frac{\varepsilon}{h}\|\widehat{\zeta} - \zeta\|) + C(h\|\delta\| + \|\theta\|). \end{aligned} \quad (3.48)$$

Here $\delta = (\delta_1, \dots, \delta_s)^T$ and $\theta = (\theta_1, \dots, \theta_s)^T$.

Proof. The essential idea is to consider the homotopy

$$\begin{pmatrix} Y_i - \eta \\ \varepsilon(Z_i - \zeta) \end{pmatrix} - h \sum_{j=1}^s a_{ij} \begin{pmatrix} f(Y_j, Z_j) \\ g(Y_j, Z_j) \end{pmatrix} = \tau \begin{pmatrix} \widehat{\eta} - \eta + h\delta_i \\ \varepsilon(\widehat{\zeta} - \zeta) + h\theta_i \end{pmatrix} \quad (3.49)$$

which relates the system (3.43) for $\tau = 0$ to the perturbed system (3.47) for $\tau = 1$. The solutions Y_i and Z_i of (3.49) are functions of τ . If we differentiate (3.49) with respect to τ and divide its second formula by h , we obtain the differential equation

$$\begin{pmatrix} I + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & M(\varepsilon/h, Y, Z) \end{pmatrix} \begin{pmatrix} \dot{Y} \\ \dot{Z} \end{pmatrix} = \begin{pmatrix} \mathbb{1} \cdot (\widehat{\eta} - \eta) + h\delta \\ (\varepsilon/h)\mathbb{1} \cdot (\widehat{\zeta} - \zeta) + \theta \end{pmatrix} \quad (3.50)$$

where $\mathbb{1} = (1, \dots, 1)^T$, $Y = (Y_1, \dots, Y_s)^T$, $Z = (Z_1, \dots, Z_s)^T$ and

$$M(\kappa, Y, Z) = \kappa I - A \otimes I \cdot \begin{pmatrix} g_z(Y_1, Z_1) & & 0 \\ & \ddots & \\ 0 & & g_z(Y_s, Z_s) \end{pmatrix}. \quad (3.51)$$

Whenever $\|Y_i - \eta\| \leq d$ and $\|Z_i - \zeta\| \leq d$ for all i , we have

$$M(\kappa, Y, Z) = \kappa I - A \otimes g_z(\eta, \zeta) + \mathcal{O}(d) \quad (3.52)$$

and it follows from (3.46) that $M^{-1}(\kappa, Y, Z)$ is uniformly bounded for $\kappa \geq 0$, if d is sufficiently small. Hence, the inverse of the matrix in (3.50) satisfies

$$\begin{pmatrix} I + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & M(\varepsilon/h, Y, Z) \end{pmatrix}^{-1} = \begin{pmatrix} I + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & \mathcal{O}(1) \end{pmatrix}$$

and the statement (3.48) follows from the fact that

$$\widehat{Y} - Y = \int_0^1 \dot{Y}(\tau) d\tau, \quad \widehat{Z} - Z = \int_0^1 \dot{Z}(\tau) d\tau. \quad \square$$

Remark 3.7. If the Runge-Kutta matrix A is only assumed to be invertible, the results of Theorems 3.5 and 3.6 still hold for $\varepsilon \leq Kh$, where K is any constant smaller than the modulus of the smallest eigenvalue of A (i.e., $K < |\lambda_{\min}|$). In this situation, the right-hand side of (3.48) is also bounded, and the same conclusions hold.

Estimation of the Remainder in the Numerical Solution

We are now in the position to estimate the remainder in (3.29). The result is the following.

Theorem 3.8 (Hairer, Lubich & Roche 1988). *Consider the stiff problem (3.1), (3.11) with initial values $y(0)$, $z(0)$ admitting a smooth solution. Apply the Runge-Kutta method (3.22)–(3.24) of classical order p and stage order q ($1 \leq q < p$). Assume that the method is A -stable, that the stability function satisfies $|R(\infty)| < 1$, and that the eigenvalues of the coefficient matrix A have positive real part. Then for any fixed constant $c > 0$ the global error satisfies for $\varepsilon \leq ch$ and $\nu \leq q + 1$*

$$\begin{aligned} y_n - y(x_n) &= \Delta y_n^0 + \varepsilon \Delta y_n^1 + \dots + \varepsilon^\nu \Delta y_n^\nu + \mathcal{O}(\varepsilon^{\nu+1}) \\ z_n - z(x_n) &= \Delta z_n^0 + \varepsilon \Delta z_n^1 + \dots + \varepsilon^\nu \Delta z_n^\nu + \mathcal{O}(\varepsilon^{\nu+1}/h). \end{aligned} \quad (3.53)$$

Here $\Delta y_n^0 = y_n^0 - y_0(x_n)$, $\Delta z_n^0 = z_n^0 - z_0(x_n)$, \dots (see Formula (3.32)) are the global errors of the method applied to the system (3.4). The estimates (3.53) hold uniformly for $h \leq h_0$ and $nh \leq \text{Const}$.

Proof. By Theorem 3.4 it suffices to prove the result for $\nu = q + 1$. We denote the truncated series of (3.25) by

$$\begin{aligned} \widehat{y}_n &= y_n^0 + \varepsilon y_n^1 + \dots + \varepsilon^\nu y_n^\nu \\ \widehat{Y}_{ni} &= Y_{ni}^0 + \varepsilon Y_{ni}^1 + \dots + \varepsilon^\nu Y_{ni}^\nu \\ \widehat{k}_{ni} &= k_{ni}^0 + \varepsilon k_{ni}^1 + \dots + \varepsilon^\nu k_{ni}^\nu \end{aligned} \quad (3.54)$$

and similarly \widehat{z}_n , \widehat{Z}_{ni} , $\widehat{\ell}_{ni}$. Further we denote

$$\Delta y_n = \widehat{y}_n - y_n, \quad \Delta Y_{ni} = \widehat{Y}_{ni} - Y_{ni}, \quad \Delta k_{ni} = \widehat{k}_{ni} - k_{ni}, \dots \quad (3.55)$$

Using (3.3) and Theorem 3.4 the statement (3.53) is then equivalent to

$$\Delta y_n = \mathcal{O}(\varepsilon^{\nu+1}), \quad \Delta z_n = \mathcal{O}(\varepsilon^{\nu+1}/h). \quad (3.56)$$

a) We first estimate the differences ΔY_{ni} , ΔZ_{ni} of the internal stages. For this we investigate the defect when (3.54) is inserted into (3.23). By our construction (3.28) it follows from (3.42) and $\nu = q + 1$ that

$$\begin{aligned}\widehat{k}_{ni} &= f(\widehat{Y}_{ni}, \widehat{Z}_{ni}) + \mathcal{O}(\varepsilon^{\nu+1}) \\ \varepsilon \widehat{\ell}_{ni} &= g(\widehat{Y}_{ni}, \widehat{Z}_{ni}) + \varepsilon^{\nu+1} \ell_{ni}^\nu + \mathcal{O}(\varepsilon^{\nu+1}).\end{aligned}\quad (3.57)$$

From (3.42) and (3.27) we know that $\ell_{ni}^\nu = \mathcal{O}(h^{-1})$. Together with (3.27) this implies

$$\begin{pmatrix} \widehat{Y}_{ni} - \widehat{y}_n \\ \varepsilon(\widehat{Z}_{ni} - \widehat{z}_n) \end{pmatrix} = h \sum_{j=1}^s a_{ij} \begin{pmatrix} f(\widehat{Y}_{nj}, \widehat{Z}_{nj}) \\ g(\widehat{Y}_{nj}, \widehat{Z}_{nj}) \end{pmatrix} + \begin{pmatrix} \mathcal{O}(h\varepsilon^{\nu+1}) \\ \mathcal{O}(\varepsilon^{\nu+1}) \end{pmatrix}, \quad (3.58)$$

which is of the form (3.47). Application of Theorem 3.6 yields

$$\begin{aligned}\|\Delta Y_{ni}\| &\leq C(\|\Delta y_n\| + \varepsilon\|\Delta z_n\|) + \mathcal{O}(\varepsilon^{\nu+1}) \\ \|\Delta Z_{ni}\| &\leq C(\|\Delta y_n\| + \frac{\varepsilon}{h}\|\Delta z_n\|) + \mathcal{O}(\varepsilon^{\nu+1}/h)\end{aligned}\quad (3.59)$$

provided that Δy_n and Δz_n are of size $\mathcal{O}(h)$. This will be justified in part (c).

b) Our next aim is to prove the recursion

$$\begin{pmatrix} \|\Delta y_{n+1}\| \\ \|\Delta z_{n+1}\| \end{pmatrix} \leq \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(\varepsilon) \\ \mathcal{O}(1) & \alpha + \mathcal{O}(\varepsilon) \end{pmatrix} \begin{pmatrix} \|\Delta y_n\| \\ \|\Delta z_n\| \end{pmatrix} + \begin{pmatrix} \mathcal{O}(\varepsilon^{\nu+1}) \\ \mathcal{O}(\varepsilon^{\nu+1}/h) \end{pmatrix} \quad (3.60)$$

where we assume again that Δy_n and Δz_n are of size $\mathcal{O}(h)$. The value $\alpha < 1$ will be given in Formula (3.63) below. The upper relation of (3.60) follows from

$$\Delta y_{n+1} = \Delta y_n + h \sum_{i=1}^s b_i \left(f(\widehat{Y}_{ni}, \widehat{Z}_{ni}) - f(Y_{ni}, Z_{ni}) \right) + \mathcal{O}(h\varepsilon^{\nu+1})$$

by the use of (3.59) and a Lipschitz condition for f .

For the verification of the second relation in (3.60) we subtract (3.57) from (3.23), and use (3.59) and (3.42) to obtain

$$\varepsilon \Delta \ell_{ni} = g_z(x_n) \Delta Z_{ni} + \mathcal{O}(\|\Delta Y_{ni}\| + h\|\Delta Z_{ni}\|) + \mathcal{O}(\varepsilon^{\nu+1}/h). \quad (3.61)$$

Here we use the notation $g_z(x) = g_z(y_0(x), z_0(x))$. Inserting $\Delta Z_{ni} = \Delta z_n + h \sum a_{ij} \Delta \ell_{nj}$ into this relation and using (3.59) again we obtain

$$\varepsilon \Delta \ell_{ni} - h \sum_{j=1}^s a_{ij} g_z(x_n) \Delta \ell_{nj} = g_z(x_n) \Delta z_n + \mathcal{O}(\|\Delta y_n\| + \varepsilon\|\Delta z_n\|) + \mathcal{O}(\varepsilon^{\nu+1}/h).$$

We now solve for $h \Delta \ell_{ni}$ and insert it into $\Delta z_{n+1} = \Delta z_n + h \sum b_i \Delta \ell_{ni}$. Since the matrix $(\varepsilon/h)I - A \otimes g_z(x_n)$ has a bounded inverse by (3.46), this gives

$$\Delta z_{n+1} = R \left(\frac{h}{\varepsilon} g_z(x_n) \right) \Delta z_n + \mathcal{O}(\|\Delta y_n\| + \varepsilon\|\Delta z_n\|) + \mathcal{O}(\varepsilon^{\nu+1}/h), \quad (3.62)$$

where $R(\mu)$ is the stability function of the method. Because of (3.11) we can apply von Neumann's theorem (Corollary IV.11.4) to estimate

$$\left\| R\left(\frac{h}{\varepsilon} g_z(x_n)\right) \right\| \leq \sup\{|R(\mu)| ; \operatorname{Re} \mu \leq -h/\varepsilon\} \leq \alpha < 1. \quad (3.63)$$

The bound α is strictly smaller than 1, because $|R(\infty)| < 1$ and $-h/\varepsilon \leq -1/c < 0$. The triangle inequality applied to (3.62) completes the proof of Formula (3.60).

c) Applying Lemma 3.9 below to the difference inequality (3.60) gives

$$\Delta y_n = \mathcal{O}(\varepsilon^{\nu+1}/h), \quad \Delta z_n = \mathcal{O}(\varepsilon^{\nu+1}/h) \quad (3.64)$$

for $nh \leq \text{Const}$. We are now in a position to justify the assumption $\Delta y_n = \mathcal{O}(h)$ and $\Delta z_n = \mathcal{O}(h)$ of the beginning of the proof. Indeed, this follows by induction on n ($\Delta y_0 = \mathcal{O}(\varepsilon^{\nu+1})$, $\Delta z_0 = \mathcal{O}(\varepsilon^{\nu+1})$) and from (3.64), because $\nu = q + 1 \geq 2$.

d) Formula (3.64) proves the desired result (3.56) for the z -component. However, the estimate (3.64) is not yet optimal for the y -component. The proof for the correct estimate is similar to that of Theorem 3.4. We have to treat more carefully the expression which gives rise to the $\mathcal{O}(\varepsilon^{\nu+1}/h)$ term in (3.61). Using (3.59) and (3.64) the same calculations which gave (3.61), now yield

$$\Delta k_{ni} = f_y(x_n) \Delta Y_{ni} + f_z(x_n) \Delta Z_{ni} + \mathcal{O}(\varepsilon^{\nu+1}) \quad (3.65a)$$

$$\varepsilon \Delta \ell_{ni} = g_y(x_n) \Delta Y_{ni} + g_z(x_n) \Delta Z_{ni} + \varepsilon^{\nu+1} \ell_{ni}^\nu + \mathcal{O}(\varepsilon^{\nu+1}). \quad (3.65b)$$

We compute ΔZ_{ni} from (3.65b) and insert it into (3.65a). This gives

$$\begin{aligned} \Delta k_{ni} - (f_z g_z^{-1})(x_n) (\varepsilon \Delta \ell_{ni} - \varepsilon^{\nu+1} \ell_{ni}^\nu) \\ = (f_y - f_z g_z^{-1} g_y)(x_n) \Delta Y_{ni} + \mathcal{O}(\varepsilon^{\nu+1}). \end{aligned} \quad (3.66)$$

Guided by this formula we put

$$\Delta u_n = \Delta y_n - (f_z g_z^{-1})(x_n) (\varepsilon \Delta z_n - \varepsilon^{\nu+1} z_n^\nu). \quad (3.67)$$

Since

$$\begin{aligned} \Delta u_{n+1} = \Delta u_n + h \sum_{i=1}^s b_i \left(\Delta k_{ni} - (f_z g_z^{-1})(x_n) (\varepsilon \Delta \ell_{ni} - \varepsilon^{\nu+1} \ell_{ni}^\nu) \right) \\ - \left((f_z g_z^{-1})(x_n + h) - (f_z g_z^{-1})(x_n) \right) (\varepsilon \Delta z_{n+1} - \varepsilon^{\nu+1} z_{n+1}^\nu) \end{aligned}$$

it follows from (3.66), (3.64), and (3.42) that

$$\|\Delta u_{n+1}\| \leq (1 + ch) \|\Delta u_n\| + \mathcal{O}(h \varepsilon^{\nu+1}). \quad (3.68)$$

As in the proof of Theorem 3.4 we deduce $\Delta u_n = \mathcal{O}(\varepsilon^{\nu+1})$ and $\Delta y_n = \mathcal{O}(\varepsilon^{\nu+1})$. \square

In the above proof we used the following result.

Lemma 3.9. *Let $\{u_n\}$, $\{v_n\}$ be two sequences of non-negative numbers satisfying (componentwise)*

$$\begin{pmatrix} u_{n+1} \\ v_{n+1} \end{pmatrix} \leq \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(\varepsilon) \\ \mathcal{O}(1) & \alpha + \mathcal{O}(\varepsilon) \end{pmatrix} \begin{pmatrix} u_n \\ v_n \end{pmatrix} + M \begin{pmatrix} h \\ 1 \end{pmatrix} \quad (3.69)$$

with $0 \leq \alpha < 1$ and $M \geq 0$. Then the following estimates hold for $\varepsilon \leq ch$, $h \leq h_0$ and $nh \leq \text{Const}$

$$\begin{aligned} u_n &\leq C(u_0 + \varepsilon v_0 + M) \\ v_n &\leq C(u_0 + (\varepsilon + \alpha^n)v_0 + M). \end{aligned} \quad (3.70)$$

Proof. We transform the matrix in (3.69) to diagonal form and so obtain

$$\begin{pmatrix} u_n \\ v_n \end{pmatrix} \leq T^{-1} \begin{pmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{pmatrix} T \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} + M \sum_{j=1}^n T^{-1} \begin{pmatrix} \lambda_1^{n-j} & 0 \\ 0 & \lambda_2^{n-j} \end{pmatrix} T \begin{pmatrix} h \\ 1 \end{pmatrix}$$

where $\lambda_1 = 1 + \mathcal{O}(h)$, $\lambda_2 = \alpha + \mathcal{O}(\varepsilon)$ are the eigenvalues and the transformation matrix T (composed of eigenvectors) satisfies

$$T = \begin{pmatrix} 1 & \mathcal{O}(\varepsilon) \\ \mathcal{O}(1) & 1 \end{pmatrix}.$$

The statement now follows from the fact that $(\alpha + \mathcal{O}(\varepsilon))^n = \mathcal{O}(\alpha^n) + \mathcal{O}(\varepsilon)$ for $\varepsilon \leq ch$ and $nh \leq \text{Const}$. \square

By combining Theorems 3.3, 3.4 and 3.8 we get the following result.

Corollary 3.10 (Hairer, Lubich & Roche 1988). *Under the assumptions of Theorem 3.8 the global error of a Runge-Kutta method satisfies*

$$y_n - y(x_n) = \mathcal{O}(h^p) + \mathcal{O}(\varepsilon h^{q+1}), \quad z_n - z(x_n) = \mathcal{O}(h^{q+1}). \quad (3.71)$$

If in addition $a_{si} = b_i$ for all i , we have

$$z_n - z(x_n) = \mathcal{O}(h^p) + \mathcal{O}(\varepsilon h^q). \quad (3.72)$$

Remarks. a) If the A -stability assumption is dropped and the coefficient matrix A is only assumed to be invertible, then the estimates of Corollary 3.10 still hold for $\varepsilon \leq Kh$ where K is a method-dependent constant (see Remark 3.7).

b) A -stability and the invertibility of the matrix A imply in general that the eigenvalues of A have positive real part. Otherwise the stability function would have to be reducible.

c) For several Runge-Kutta methods satisfying $a_{si} = b_i$ the estimate (3.71) for the y -component can be improved. E.g., for Radau IIA and for Lobatto IIIC one has $y_n - y(x_n) = \mathcal{O}(h^p) + \mathcal{O}(\varepsilon^2 h^q)$. This follows from Table VII.4.1 below.

d) A completely different proof of the estimates (3.71) is given by Nipp & Stoffer (1995). They show that the Runge-Kutta method, considered as a discrete

dynamical system, admits an attractive invariant manifold $M_{h,\varepsilon}$, which is close to the invariant manifold M_ε of the problem (3.1). Studying the closeness of the two manifolds, they obtain the error estimates (3.71) without considering ε -expansions.

e) The analogues of Theorem 3.8 and Corollary 3.10 for Rosenbrock methods are given in Hairer, Lubich & Roche (1989).

f) Estimates for $p = q$ are given in Exercise 3 below.

Numerical Confirmation

The estimates of Corollary 3.10 can be observed numerically. As an example of (3.1) we choose the van der Pol equation

$$\begin{aligned} y' &= z \\ \varepsilon z' &= (1 - y^2)z - y \end{aligned} \quad (3.73)$$

with $\varepsilon = 10^{-5}$ and initial values

$$y(0) = 2, \quad z(0) = -0.6666654321121172 \quad (3.74)$$

on the smooth solution (Exercise 2).

Table 3.1 shows the methods of our experiment together with the theoretical error bounds. In Fig. 3.1 we have plotted the relative global error at $x_{end} = 0.5$ as a function of the step size h , which was taken constant over the considered interval. The use of logarithmic scales in both directions makes the curves appear as straight lines of slope r , whenever the leading term of the global error behaves like $Const \cdot h^r$. The figures show complete agreement with our theoretical results.

Table 3.1. Global errors predicted by Corollary 3.10

Method	$a_{si} = b_i$	y -comp.	z -comp.
Radau IA	no	$h^{2s-1} + \varepsilon h^s$	h^s
Radau IIA	yes	$h^{2s-1} + \varepsilon^2 h^s$	$h^{2s-1} + \varepsilon h^s$
Lobatto IIIC	yes	$h^{2s-2} + \varepsilon^2 h^{s-1}$	$h^{2s-2} + \varepsilon h^{s-1}$
SDIRK (IV.6.16)	yes	$h^4 + \varepsilon h^2$	$h^4 + \varepsilon h$
SDIRK (IV.6.18)	no	$h^4 + \varepsilon h^2$	h^2

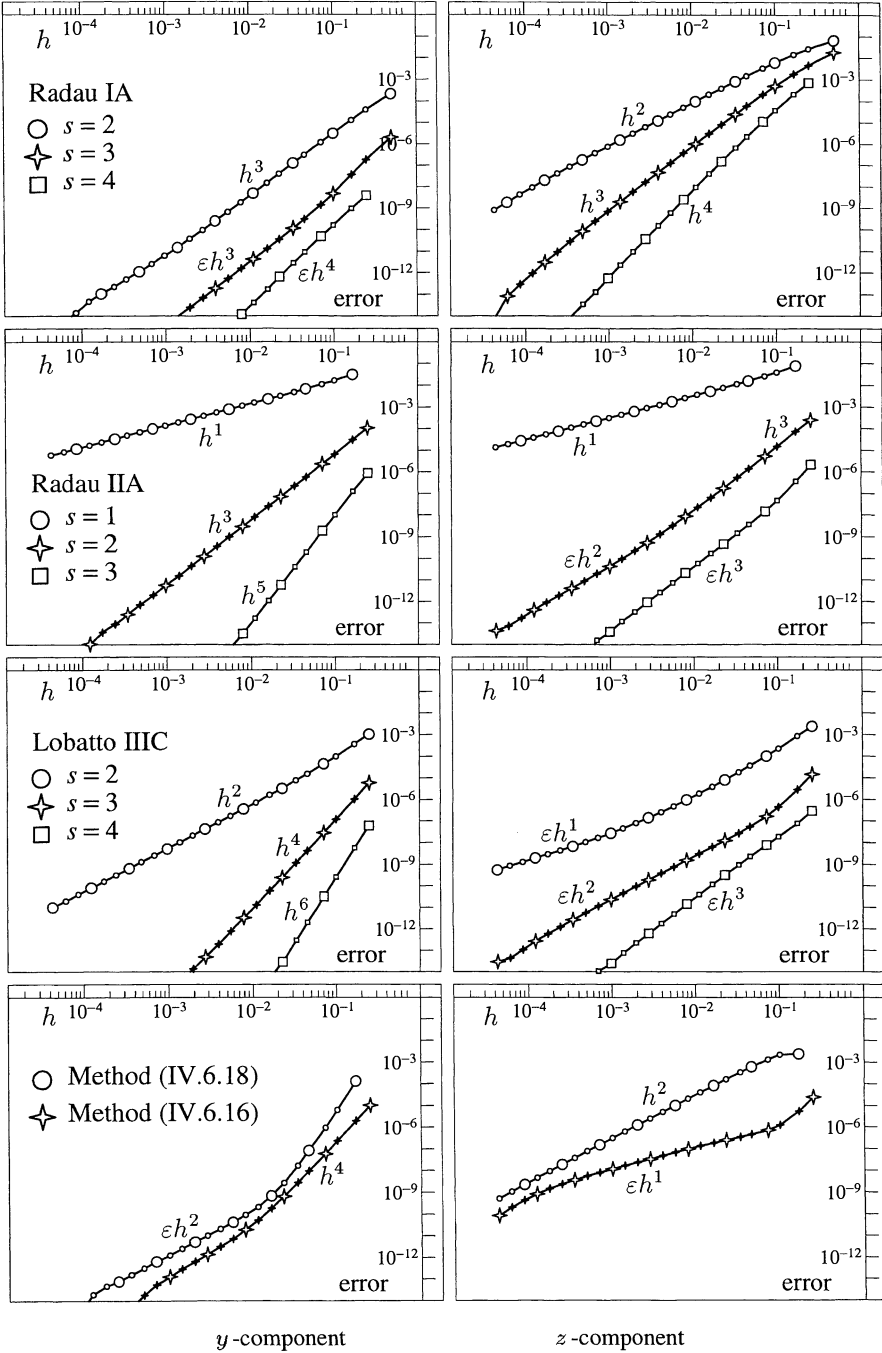


Fig. 3.1. Global error versus the step size

Perturbed Initial Values

When integrating a singular perturbation problem, the numerical solution approximates the smooth solution only within the given tolerance Tol . It is therefore interesting to investigate the influence of perturbations in the initial values on the global and local errors of the method. Let us begin with a numerical experiment. We perturb the $z(0)$ value of (3.74) by an amount of 10^{-6} and apply the Radau IIA methods to the problem (3.73). For the global error at $x_{end}=0.5$ we obtain exactly the same results as in Fig. 3.1. This shows that the perturbation is completely damped out during integration. The results for the local error show a different behaviour and are displayed in Fig. 3.2. We observe the presence of a “hump”, exactly as in Fig. IV.7.4 and in Fig. IV.8.2.

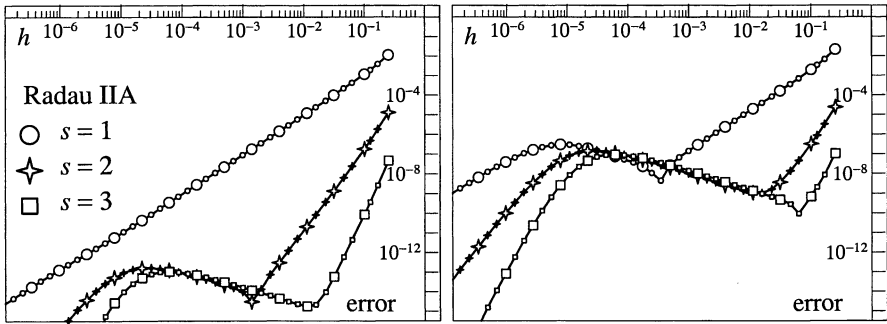


Fig. 3.2. Local error of Radau IIA (perturbed initial value)

In order to explain this phenomenon we denote by (y_0, z_0) the considered initial value, and by (y_1, z_1) the numerical solution after one step with step size h . The exact solution $y(x), z(x)$ passing through (y_0, z_0) will have a boundary layer, and (under suitable assumptions, see Theorem 3.2) can be written as

$$y(x) = \tilde{y}(x) + \mathcal{O}(\varepsilon e^{-x/\varepsilon}), \quad z(x) = \tilde{z}(x) + \mathcal{O}(e^{-x/\varepsilon}). \quad (3.75)$$

Here $\tilde{y}(x), \tilde{z}(x)$ represents a smooth solution of (3.1). We denote by $\tilde{y}_0 = \tilde{y}(0)$, $\tilde{z}_0 = \tilde{z}(0)$ the initial values on this smooth solution, and by $(\tilde{y}_1, \tilde{z}_1)$ the numerical approximation obtained by the same method with step size h and initial values $(\tilde{y}_0, \tilde{z}_0)$. The local error can now be written as

$$z_1 - z(h) = (z_1 - \tilde{z}_1) + (\tilde{z}_1 - \tilde{z}(h)) + (\tilde{z}(h) - z(h)) \quad (3.76)$$

and similarly for the y -component. The last term in (3.76), which is of size $\mathcal{O}(Tol \cdot e^{-h/\varepsilon})$, can be neglected if the step size h is significantly larger than ε . The term $\tilde{z}_1 - \tilde{z}(h)$ represents the local error in the “smooth” situation and is bounded by at least $\mathcal{O}(h^{q+1})$ (apply Corollary 3.10 with $n = 1$). It can be observed in Fig. 3.2 whenever h or the error is large. The difference $z_1 - \tilde{z}_1$ is the term which causes the irregularity in Fig. 3.2. Using Theorem 3.6 (with $\delta = 0$,

$\theta = 0$, $\hat{\eta} - \eta = \mathcal{O}(\varepsilon \cdot Tol)$, $\hat{\zeta} - \zeta = \mathcal{O}(Tol)$) and the ideas of the proof of Theorem 3.8 (in particular Eq. (3.62)) we obtain

$$\begin{aligned} z_1 - \tilde{z}_1 &= R\left(\frac{h}{\varepsilon} g_z(0)\right)(z_0 - \tilde{z}_0) + \mathcal{O}(\varepsilon \cdot Tol) \\ y_1 - \tilde{y}_1 &= \mathcal{O}(\varepsilon \cdot Tol). \end{aligned} \quad (3.77)$$

For $\varepsilon < h$ we develop

$$R\left(\frac{h}{\varepsilon} g_z(0)\right) = R(\infty) + C \frac{\varepsilon}{h} g_z^{-1}(0) + \mathcal{O}\left(\left(\frac{\varepsilon}{h}\right)^2\right). \quad (3.78)$$

This shows that an h -independent expression $R(\infty)(z_0 - \tilde{z}_0) = \mathcal{O}(Tol)$ will be observed in the local error, if $R(\infty) \neq 0$. For methods with $R(\infty) = 0$ (such as Radau IIA) the dominant part in $z_1 - \tilde{z}_1$ is $C(\varepsilon/h)g_z^{-1}(0)(z_0 - \tilde{z}_0) = \mathcal{O}(Tol \cdot \varepsilon/h)$. This term can be observed in Fig. 3.2 as a straight line of slope -1 . Thus in this region the local error increases like h^{-1} when h decreases. A similar perturbation, multiplied however by ε , is observed for the y -component.

This is not a serious drawback for a numerical implementation, because the phenomenon appears only for step sizes where the local error is smaller than Tol .

Exercises

1. Prove that the statement of Theorem 3.2 remains valid, if the assumption (3.11) is replaced by

the eigenvalues λ of $g_z(y, z)$ satisfy $\operatorname{Re} \lambda \leq -1$

for all y, z in a neighbourhood of the solution $y_0(x)$, $z_0(x)$ of the reduced system.

Hint. Split the interval into a finite number of small subintervals and construct for each of them an inner product norm such that, after a rescaling of ε , (3.11) holds (see Nevanlinna 1976).

2. Let $y(0) = 2$; find the corresponding $z(0)$ for the van der Pol equation (3.73), such that its solution is smooth.

Result.

$$z(0) = -\frac{2}{3} + \frac{10}{81}\varepsilon - \frac{292}{2187}\varepsilon^2 - \frac{1814}{19683}\varepsilon^3 + \mathcal{O}(\varepsilon^4).$$

3. If the assumption $q < p$ (p classical order, q stage order) is dropped in Corollary 3.10, we still have

$$y_n - y(x_n) = \mathcal{O}(h^p), \quad z_n - z(x_n) = \mathcal{O}(h^p).$$

Prove this statement. The implicit Euler method and the SIRK methods of Lemma IV.8.1 are typical examples with $p = q$.

Hint. Apply Corollary 3.10 with q reduced by 1.

VI.4 Rosenbrock Methods

This section is devoted to the extension of Rosenbrock methods (see Sect. IV.7) to differential-algebraic equations in semi-explicit form

$$y' = f(y, z), \quad y(x_0) = y_0 \quad (4.1a)$$

$$0 = g(y, z), \quad z(x_0) = z_0. \quad (4.1b)$$

We suppose that g_z is invertible (see (1.7)), so that the problem is of index 1. We shall obtain new methods for the numerical solution of such problems, and at the same time get more insight into the behaviour of Rosenbrock methods for stiff differential equations. In particular, the phenomenon of Fig. IV.7.4 will be explained.

Definition of the Method

The main advantage of Rosenbrock methods over implicit Runge-Kutta methods is that nonlinear systems are completely avoided. The state space form method (transforming (4.1) to $y' = f(y, G(y))$) would destroy this advantage. This is one more reason for considering the ε -embedding method. For the problem (1.5) a Rosenbrock method reads

$$\begin{pmatrix} k_i \\ \ell_i \end{pmatrix} = h \begin{pmatrix} f(v_i, w_i) \\ \varepsilon^{-1} g(v_i, w_i) \end{pmatrix} + h \begin{pmatrix} f_y & f_z \\ \varepsilon^{-1} g_y & \varepsilon^{-1} g_z \end{pmatrix} (y_0, z_0) \sum_{j=1}^i \gamma_{ij} \begin{pmatrix} k_j \\ \ell_j \end{pmatrix} \quad (4.2)$$

$$\begin{pmatrix} v_i \\ w_i \end{pmatrix} = \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} + \sum_{j=1}^{i-1} \alpha_{ij} \begin{pmatrix} k_j \\ \ell_j \end{pmatrix}, \quad \begin{pmatrix} y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} + \sum_{i=1}^s b_i \begin{pmatrix} k_i \\ \ell_i \end{pmatrix}. \quad (4.3a)$$

If we multiply the second line of (4.2) by ε and then put $\varepsilon = 0$ we obtain

$$\begin{pmatrix} k_i \\ 0 \end{pmatrix} = h \begin{pmatrix} f(v_i, w_i) \\ g(v_i, w_i) \end{pmatrix} + h \begin{pmatrix} f_y & f_z \\ g_y & g_z \end{pmatrix} (y_0, z_0) \sum_{j=1}^i \gamma_{ij} \begin{pmatrix} k_j \\ \ell_j \end{pmatrix}. \quad (4.3b)$$

Formulas (4.3a) and (4.3b) together constitute the extension of a Rosenbrock method to the problem (4.1). This type of method was first considered by Michelsen (1976) (quoted by Feng, Holland & Gallun (1984)). Further studies are due to Roche

(1988). We remark that the computation of (k_i, ℓ_i) from (4.3b) requires the solution of a linear system with matrix

$$\begin{pmatrix} I - \gamma h f_y & -\gamma h f_z \\ -\gamma h g_y & -\gamma h g_z \end{pmatrix} \quad (4.4)$$

where all derivatives are evaluated at (y_0, z_0) . For nonsingular g_z , nonzero γ , and small enough $h > 0$, this matrix is invertible. This can be seen by dividing the lower blocks by γh and then putting $h = 0$.

Non-autonomous equations. If the functions f and g in (4.1) also depend on x , we replace (4.3b) by

$$\begin{pmatrix} k_i \\ 0 \end{pmatrix} = h \begin{pmatrix} f(x_0 + \alpha_i h, v_i, w_i) \\ g(x_0 + \alpha_i h, v_i, w_i) \end{pmatrix} + h \begin{pmatrix} f_y & f_z \\ g_y & g_z \end{pmatrix} \sum_{j=1}^i \gamma_{ij} \begin{pmatrix} k_j \\ \ell_j \end{pmatrix} + h^2 \gamma_i \begin{pmatrix} f_x \\ g_x \end{pmatrix} \quad (4.5)$$

(compare with (IV.7.4a) and recall the definition of α_i and γ_i in (IV.7.5)). All derivatives are evaluated at the initial value (x_0, y_0, z_0) .

Problems of the form $Mu' = \varphi(u)$. Rosenbrock formulas for these problems have been developed in Sect. IV.7 (Formula (IV.7.4b)) in the case of regular M . This formula is also applicable for singular M , and can be justified as follows: It is theoretically possible to apply the transformation (1.20) so that M becomes the block-diagonal matrix with entries I and 0 . The method (IV.7.4b) is then identical to method (4.3). Therefore, the theory to be developed in this section will also be valid for Rosenbrock method (IV.7.4b) applied to index 1 problems of the form $Mu' = \varphi(u)$.

Having introduced a new class of methods, we must study their order conditions. As usual, this is done by Taylor expansion of both the exact and the numerical solution (similar to Section II.2). A nice correspondence between the order conditions and certain rooted trees with two different kinds of vertices will be obtained (Roche 1988).

Derivatives of the Exact Solution

In contrast to Sect. II.2, where we used “hordes of indices” (see Dieudonné’s preface to his “Foundations of Modern Analysis”) to show us the way through the “woud met bomen” (Hundsdofer), we here write higher derivatives as multilinear mappings. For example, the expression

$$\sum_{j,k} \frac{\partial^2 g_i}{\partial y_j \partial z_k} \cdot u_j v_k \quad \text{is written as} \quad g_{yz}(u, v),$$

which simplifies the subsequent formulas.

We differentiate (4.1b) to obtain $0 = g_y \cdot y' + g_z \cdot z'$ and, equivalently,

$$z' = (-g_z^{-1})g_y f. \quad (4.6)$$

We now differentiate successively (4.1a) and (4.6) with respect to x . We use the formula

$$(-g_z^{-1})'u = (-g_z^{-1})\left(g_{zy}((-g_z^{-1})u, f) + g_{zz}((-g_z^{-1})u, (-g_z^{-1})g_y f)\right) \quad (4.7)$$

which is a consequence of $(A^{-1}(x))' = -A^{-1}(x)A'(x)A^{-1}(x)$ and the chain rule. This gives

$$y'' = f_y \cdot y' + f_z \cdot z' = f_y f + f_z (-g_z^{-1})g_y f \quad (4.8)$$

$$\begin{aligned} z'' = & (-g_z^{-1})\left(g_{zy}((-g_z^{-1})g_y f, f) + g_{zz}((-g_z^{-1})g_y f, (-g_z^{-1})g_y f)\right) \\ & + (-g_z^{-1})\left(g_{yy}(f, f) + g_{yz}(f, (-g_z^{-1})g_y f)\right) \\ & + (-g_z^{-1})g_y\left(f_y f + f_z (-g_z^{-1})g_y f\right). \end{aligned} \quad (4.9)$$

Clearly, these expressions soon become very complicated and a graphical representation of the terms in (4.8) and (4.9) is desirable.

Trees and Elementary Differentials

We shall identify each occurring f with a meagre vertex, and each of its derivatives with an upward leaving branch. The expression $(-g_z^{-1})g$ is identified with a fat vertex. The derivatives of g therein are again indicated by upwards leaving branches. For example, the second expression of (4.8) and the first one of (4.9) correspond to the trees in Fig. 4.1.

The above formulas for y', z', y'', z'' thus become

$$\begin{array}{ll} y' = \bullet & z' = \text{fat vertex with one upward branch} \\ y'' = \text{meagre vertex with two upward branches} & z'' = \text{fat vertex with two upward branches, each having two upward branches} \end{array} \quad (4.10)$$

The first and fourth expressions in (4.9) are identical, because $g_{zy}(u, v) = g_{yz}(v, u)$. This is in nice accordance with the fact that the corresponding trees are topologically equivalent. The lowest vertex of a tree will be called its *root*.

We see that derivatives of y are characterized by trees with a *meagre root*. These trees will be denoted by t or t_i , the tree consisting only of the root (for y') being τ_y . Derivatives of z have trees with a *fat root*. These will be written as u or u_i , the tree for z' being τ_z .

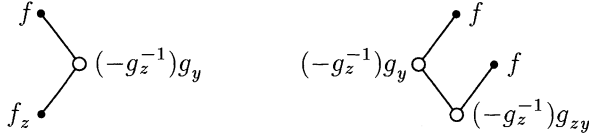


Fig. 4.1. Graphical representation of elementary differentials

Definition 4.1. Let $DAT = DAT_y \cup DAT_z$ denote the set of (differential algebraic rooted) *trees* defined recursively by

- a) $\tau_y \in DAT_y, \tau_z \in DAT_z$;
- b) $[t_1, \dots, t_m, u_1, \dots, u_n]_y \in DAT_y$
if $t_1, \dots, t_m \in DAT_y$ and $u_1, \dots, u_n \in DAT_z$;
- c) $[t_1, \dots, t_m, u_1, \dots, u_n]_z \in DAT_z$
if $t_1, \dots, t_m \in DAT_y, u_1, \dots, u_n \in DAT_z$, and $(m, n) \neq (0, 1)$.

Here $[t_1, \dots, t_m, u_1, \dots, u_n]_y$ and $[t_1, \dots, t_m, u_1, \dots, u_n]_z$ represent unordered $(m+n)$ -tuples.

The graphical representation of these trees is as follows: if we connect the roots of $t_1, \dots, t_m, u_1, \dots, u_n$ by $m+n$ branches to a new meagre vertex (the new root) we obtain $[t_1, \dots, t_m, u_1, \dots, u_n]_y$; if we connect them to a new fat vertex we obtain $[t_1, \dots, t_m, u_1, \dots, u_n]_z$. For example, the two trees of Fig. 4.1 can be written as $[\tau_z]_y$ and $[\tau_z, \tau_y]_z$.

Definition 4.2. The *order* of a tree $t \in DAT_y$ or $u \in DAT_z$, denoted by $\varrho(t)$ or $\varrho(u)$, is the number of its meagre vertices.

We see in (4.10) that this definition of order coincides with the derivative order of $y^{(i)}$ or $z^{(i)}$ as far as they are computed there.

We next give a recursive definition of the one-to-one correspondence between the trees in (4.10) and the expressions in (4.8) and (4.9).

Definition 4.3. The *elementary differentials* $F(t)$ (or $F(u)$) corresponding to trees in DAT are defined as follows:

- a) $F(\tau_y) = f, \quad F(\tau_z) = (-g_z^{-1})g_y f$,
- b) $F(t) = \frac{\partial^{m+n} f}{\partial y^m \partial z^n} \left(F(t_1), \dots, F(t_m), F(u_1), \dots, F(u_n) \right)$
if $t = [t_1, \dots, t_m, u_1, \dots, u_n]_y \in DAT_y$,
- c) $F(u) = (-g_z)^{-1} \frac{\partial^{m+n} g}{\partial y^m \partial z^n} \left(F(t_1), \dots, F(t_m), F(u_1), \dots, F(u_n) \right)$
if $u = [t_1, \dots, t_m, u_1, \dots, u_n]_z \in DAT_z$.

Because of the symmetry of partial derivatives, this definition is unaffected by a permutation of $t_1, \dots, t_m, u_1, \dots, u_n$ and therefore the functions $F(t)$ and $F(u)$ are well defined.

Taylor Expansion of the Exact Solution

In order to get more insight into the process of (4.8) and (4.9) we study the differentiation of an elementary differential with respect to x . By Leibniz' rule the differentiation of $F(t)$ (or $F(u)$) gives a sum of new elementary differentials which are obtained by the following four rules:

- i) attach to each vertex a branch with τ_y (derivative of f or g with respect to y and addition of the factor $y' = f$);
- ii) attach to each vertex a branch with τ_z (derivative of f or g with respect to z and addition of the factor $z' = (-g_z^{-1})g_y f$);
- iii) split each fat vertex into two new fat vertices (linked by a new branch) and attach to the lower of these fat vertices a branch with τ_y ;
- iv) as in (iii) split each fat vertex into two new fat vertices, but attach this time to the lower of the new fat vertices a branch with τ_z .

The rules (iii) and (iv) correspond to the differentiation of $(-g_z^{-1})$ and follow at once from (4.7). We observe that the differentiation of a tree of order q (or, more precisely, of its corresponding elementary differential) generates trees of order $q + 1$.

As was the case in Sect. II.2, some of these trees appear *several times* in the derivative (as the first and fourth tree for z'' in (4.10)). In order to distinguish all these trees, we indicate the *order of generation* of the meagre vertices by labels. This is demonstrated, for the first derivatives of y , in Fig. 4.2. Since in the above differentiation process the new meagre vertex is always an end-vertex of the tree, the labelling thus obtained is necessarily increasing from the root upwards along each branch.

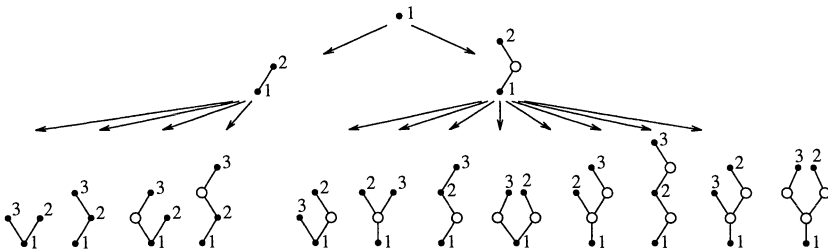


Fig. 4.2. Monotonically labelled trees ($LDAT_y$)

Definition 4.4. A tree $t \in DAT_y$ (or $u \in DAT_z$) together with a monotonic labelling of its meagre vertices is called a *monotonically labelled tree*. The sets of all such monotonically labelled trees are denoted by $LDAT_y$, $LDAT_z$ and $LDAT$.

Definition 4.2 (order of a tree) and Definition 4.3 (elementary differential) are extended in a natural way to monotonically labelled trees. We can therefore write the derivatives of the exact solution as follows:

Theorem 4.5 (Roche 1988). *For the exact solution of (4.1) we have:*

$$\begin{aligned} y^{(q)}(x_0) &= \sum_{t \in LDAT_y, \varrho(t)=q} F(t)(y_0, z_0) = \sum_{t \in DAT_y, \varrho(t)=q} \alpha(t) F(t)(y_0, z_0) \\ z^{(q)}(x_0) &= \sum_{u \in LDAT_z, \varrho(u)=q} F(u)(y_0, z_0) = \sum_{u \in DAT_z, \varrho(u)=q} \alpha(u) F(u)(y_0, z_0). \end{aligned}$$

The integer coefficients $\alpha(t)$ and $\alpha(u)$ indicate the number of possible monotonic labellings of a tree.

Proof. For $q = 1$ and $q = 2$ this is just (4.1a), (4.6), (4.8) and (4.9). For general q the above differentiation process of trees generates all elements of $LDAT$, each element exactly once. If the sum is taken over DAT_y and DAT_z , the factors $\alpha(t)$ and $\alpha(u)$ must be added. \square

Taylor Expansion of the Numerical Solution

Our next aim is to prove an analogue of Theorem 4.5 for the numerical solution of a Rosenbrock method. We consider y_1, z_1 as functions of the step size h and compute their derivatives. From (4.3a) it follows that

$$y_1^{(q)}(0) = \sum_{i=1}^s b_i k_i^{(q)}(0), \quad z_1^{(q)}(0) = \sum_{i=1}^s b_i \ell_i^{(q)}(0). \quad (4.11)$$

Consequently we have to compute the derivatives of k_i and ℓ_i . This is done as for Runge-Kutta methods (Sect. II.2) or for Rosenbrock methods applied to ordinary differential equations (Sect. IV.7).

We differentiate the first line of (4.3b) with respect to h . Using Leibniz' rule (II.2.4) this yields for $h = 0$

$$k_i^{(q)} = q(f(v_i, w_i))^{(q-1)} + (f_y)_0 q \sum_{j=1}^i \gamma_{ij} k_j^{(q-1)} + (f_z)_0 q \sum_{j=1}^i \gamma_{ij} \ell_j^{(q-1)}. \quad (4.12)$$

The index 0 in $(f_y)_0$ and $(f_z)_0$ indicates that the derivatives are evaluated at (y_0, z_0) . The second line of (4.3b) is divided by h before differentiation. This gives (again for $h = 0$)

$$0 = (g(v_i, w_i))^{(q)} + (g_y)_0 \sum_{j=1}^i \gamma_{ij} k_j^{(q)} + (g_z)_0 \sum_{j=1}^i \gamma_{ij} \ell_j^{(q)}. \quad (4.13)$$

The derivatives of f and g can be computed by Faà di Bruno's formula (Lemma II.2.8). This yields

$$(f(v_i, w_i))^{(q-1)} = \sum \frac{\partial^{m+n} f(v_i, w_i)}{\partial y^m \partial z^n} (v_i^{(\mu_1)}, \dots, v_i^{(\mu_m)}, w_i^{(\nu_1)}, \dots, w_i^{(\nu_n)}) \quad (4.14)$$

where the sum is over all “special $LDAT_y$ ’s” of order q . These are monotonically labelled trees $[t_1, \dots, t_m, u_1, \dots, u_n]_y$ where t_j and u_j do not have any ramification and all their vertices are meagre with the exception of the roots of u_1, \dots, u_n . The integers μ_j and ν_j are the orders of t_j and u_j , respectively. They satisfy $\mu_1 + \dots + \mu_m + \nu_1 + \dots + \nu_n = q - 1$. Similarly we apply Faà di Bruno’s formula to g and obtain

$$(g(v_i, w_i))^{(q)} = \sum \frac{\partial^{m+n} g(v_i, w_i)}{\partial y^m \partial z^n} (v_i^{(\mu_1)}, \dots, v_i^{(\mu_m)}, w_i^{(\nu_1)}, \dots, w_i^{(\nu_n)}) + g_z(v_i, w_i) w_i^{(q)}. \quad (4.15)$$

Here the sum is over all “special $LDAT_z$ ’s” of order q . They are defined as above but have a fat vertex. The integers μ_j, ν_j satisfy $\mu_1 + \dots + \mu_m + \nu_1 + \dots + \nu_n = q$. The term with g_z is written separately, because (by the definition of $LDAT_z$) $[u_1]_z$ is not an admissible tree.

We are now in a position to compute the derivatives of k_i and ℓ_i . For this it is convenient to introduce the notation

$$\beta_{ij} = \alpha_{ij} + \gamma_{ij} \quad (4.16)$$

(with $\alpha_{ii} = 0$) as in (IV.7.12). We also need the inverse of the matrix (β_{ij}) , whose coefficients we denote by ω_{ij} :

$$(\omega_{ij}) = (\beta_{ij})^{-1}. \quad (4.17)$$

Theorem 4.6. *The derivatives of k_i and ℓ_i satisfy*

$$\begin{aligned} k_i^{(q)}(0) &= \sum_{t \in LDAT_y, \varrho(t)=q} \gamma(t) \Phi_i(t) F(t)(y_0, z_0) \\ \ell_i^{(q)}(0) &= \sum_{u \in LDAT_z, \varrho(u)=q} \gamma(u) \Phi_i(u) F(u)(y_0, z_0), \end{aligned} \quad (4.18)$$

where the coefficients $\Phi_i(t)$ and $\Phi_i(u)$ are given by $\Phi_i(\tau_y) = 1$, $\Phi_i(\tau_z) = 1$ and

$$\Phi_i(t) = \begin{cases} \sum_{\mu_1, \dots, \mu_m, \nu_1, \dots, \nu_n} \alpha_{i\mu_1} \cdots \alpha_{i\mu_m} \alpha_{i\nu_1} \cdots \alpha_{i\nu_n} \cdot \Phi_{\mu_1}(t_1) \cdots \Phi_{\mu_m}(t_m) \Phi_{\nu_1}(u_1) \cdots \Phi_{\nu_n}(u_n) & \text{if } t = [t_1, \dots, t_m, u_1, \dots, u_n]_y \text{ and } m+n \geq 2 \\ \sum_j \beta_{ij} \Phi_j(t_1) & \text{if } t = [t_1]_y \\ \sum_j \beta_{ij} \Phi_j(u_1) & \text{if } t = [u_1]_y, \end{cases}$$

$$\Phi_i(u) = \begin{cases} \sum_{j, \mu_1, \dots, \mu_m, \nu_1, \dots, \nu_n} \omega_{ij} \alpha_{j\mu_1} \cdots \alpha_{j\mu_m} \alpha_{j\nu_1} \cdots \alpha_{j\nu_n} \\ \quad \cdot \Phi_{\mu_1}(t_1) \cdots \Phi_{\mu_m}(t_m) \Phi_{\nu_1}(u_1) \cdots \Phi_{\nu_n}(u_n) \\ \quad \text{if } u = [t_1, \dots, t_m, u_1, \dots, u_n]_z \text{ and } m+n \geq 2 \\ \Phi_i(t_1) \quad \text{if } u = [t_1]_z \end{cases}$$

and the integer coefficients $\gamma(t)$ and $\gamma(u)$ are defined by $\gamma(\tau_y) = 1$, $\gamma(\tau_z) = 1$ and

$$\begin{aligned} \gamma(t) &= \varrho(t) \gamma(t_1) \cdots \gamma(t_m) \gamma(u_1) \cdots \gamma(u_n) & \text{if } t = [t_1, \dots, t_m, u_1, \dots, u_n]_y \\ \gamma(u) &= \gamma(t_1) \cdots \gamma(t_m) \gamma(u_1) \cdots \gamma(u_n) & \text{if } u = [t_1, \dots, t_m, u_1, \dots, u_n]_z. \end{aligned}$$

Proof. By (4.3a) we have

$$v_i^{(\mu)} = \sum_{j=1}^{i-1} \alpha_{ij} k_j^{(\mu)}, \quad w_i^{(\nu)} = \sum_{j=1}^{i-1} \alpha_{ij} \ell_j^{(\nu)}. \quad (4.19)$$

We now insert (4.19) into (4.14) and the resulting formula for $(f(v_i, w_i))^{(q-1)}$ into (4.12). This yields (all expressions have to be evaluated at $h = 0$)

$$\begin{aligned} k_i^{(q)} &= q \sum_{m+n \geq 2} \frac{\partial^{m+n} f(y_0, z_0)}{\partial y^m \partial z^n} \left(\sum_{j=1}^{i-1} \alpha_{ij} k_j^{(\mu_1)}, \dots, \sum_{j=1}^{i-1} \alpha_{ij} \ell_j^{(\nu_1)}, \dots \right) \\ &\quad + q(f_y)_0 \sum_{j=1}^i \beta_{ij} k_j^{(q-1)} + q(f_z)_0 \sum_{j=1}^i \beta_{ij} \ell_j^{(q-1)}. \end{aligned} \quad (4.20)$$

The same analysis for the second component leads to

$$\begin{aligned} 0 &= \sum_{m+n \geq 2} \frac{\partial^{m+n} g(y_0, z_0)}{\partial y^m \partial z^n} \left(\sum_{j=1}^{i-1} \alpha_{ij} k_j^{(\mu_1)}, \dots, \sum_{j=1}^{i-1} \alpha_{ij} \ell_j^{(\nu_1)}, \dots \right) \\ &\quad + (g_y)_0 \sum_{j=1}^i \beta_{ij} k_j^{(q)} + (g_z)_0 \sum_{j=1}^i \beta_{ij} \ell_j^{(q)}. \end{aligned} \quad (4.21)$$

The sums in (4.20) and (4.21) are over elements of $LDAT$ exactly as in (4.14) and (4.15). Equation (4.21) allows us to extract $\ell_i^{(q)}$ if we use the inverse of (β_{ij}) . This gives

$$\begin{aligned} \ell_i^{(q)} &= (-g_z)_0^{-1} \sum_{j=1}^i \omega_{ij} \sum_{m+n \geq 2} \frac{\partial^{m+n} g(y_0, z_0)}{\partial y^m \partial z^n} \left(\sum_{\kappa=1}^{j-1} \alpha_{j\kappa} k_{\kappa}^{(\mu_1)}, \dots, \sum_{\kappa=1}^{j-1} \alpha_{j\kappa} \ell_{\kappa}^{(\nu_1)}, \dots \right) \\ &\quad + ((-g_z^{-1})g_y)_0 k_i^{(q)}. \end{aligned} \quad (4.22)$$

The proof of Formula (4.18) is now by induction on q . The case $q = 1$ follows immediately from (4.12) and (4.13). For general q , we insert the induction hypothesis into (4.20) and (4.22), exploit the multilinearity of the derivatives, and arrange the summations as in the proof of Theorem II.2.11. \square

Finally, Eq. (4.11) yields the derivatives of the numerical solution.

Theorem 4.7 (Roche 1988). *The numerical solution of (4.3) satisfies:*

$$\begin{aligned} y_1^{(q)}|_{h=0} &= \sum_{t \in LDAT_y, \varrho(t)=q} \gamma(t) \sum_{i=1}^s b_i \Phi_i(t) F(t)(y_0, z_0) \\ z_1^{(q)}|_{h=0} &= \sum_{u \in LDAT_z, \varrho(u)=q} \gamma(u) \sum_{i=1}^s b_i \Phi_i(u) F(u)(y_0, z_0) \end{aligned}$$

where the coefficients γ and Φ_i are given in Theorem 4.6. □

Order Conditions

Comparing Theorem 4.5 and 4.7 we obtain

Theorem 4.8. *For the Rosenbrock method (4.3) we have:*

$$\begin{aligned} y(x_0 + h) - y_1 &= \mathcal{O}(h^{p+1}) \quad \text{iff} \\ \sum_{i=1}^s b_i \Phi_i(t) &= \frac{1}{\gamma(t)} \quad \text{for } t \in DAT_y, \varrho(t) \leq p; \\ z(x_0 + h) - z_1 &= \mathcal{O}(h^{q+1}) \quad \text{iff} \\ \sum_{i=1}^s b_i \Phi_i(u) &= \frac{1}{\gamma(u)} \quad \text{for } u \in DAT_z, \varrho(u) \leq q, \end{aligned}$$

where the coefficients Φ_i and γ are those of Theorem 4.6. □

Repeated application of the recursive definition of Φ_i in Theorem 4.6 yields the following algorithm:

Forming the Order Condition for a Given Tree: attach to each meagre vertex one summation index, and to each fat vertex two indices (one above the other). Then the left hand side of the order condition is a sum over all indices of a product with factors

- b_i if “ i ” is the index of the root (the lower index if the root is fat);
- α_{ij} if “ j ” lies directly above “ i ” and “ i ” is multiply branched;
- β_{ij} if “ j ” lies directly above “ i ” and “ i ” is singly branched;
- ω_{ij} if “ i, j ” are the two indices of a fat vertex (“ i ” below “ j ”).

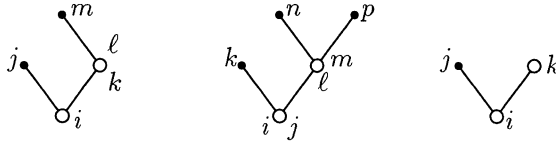


Fig. 4.3. Trees with labelling

As an example, we present the order conditions for the first two trees of Fig. 4.3.

$$\sum_{i,j,k,\ell,m} b_i \alpha_{ij} \alpha_{ik} \omega_{k\ell} \beta_{\ell m} = \frac{1}{3} \quad (4.23)$$

$$\sum_{i,j,k,\ell,m,n,p} b_i \omega_{ij} \alpha_{jk} \alpha_{j\ell} \omega_{\ell m} \alpha_{mn} \alpha_{mp} = 1. \quad (4.24)$$

The condition (4.23) can be further simplified if we use the fact that (ω_{ij}) is the inverse of the matrix (β_{ij}) . Indeed, (4.23) is equivalent to

$$\sum_{i,j,k} b_i \alpha_{ij} \alpha_{ik} = \frac{1}{3}$$

which is the order condition for the third tree in Fig. 4.3. Exploiting this reduction systematically we arrive at the following result.

Lemma 4.9. *For a Rosenbrock method (4.3) the order conditions corresponding to one of the following situations are redundant:*

- a) a fat vertex is singly branched.
- b) a singly branched vertex is followed by a fat vertex. □

The subset of DAT_y which consists of trees with only meagre vertices, is simply T (the set of trees of Sect. II.2). The corresponding order conditions are those given in Sect. IV.7. Consequently, a p -th order Rosenbrock method has to satisfy all “classical” order conditions and, in addition, several “algebraic” conditions. The first of these new order conditions are given in Table 4.1. We have included the polynomial $p_t(\gamma)$ in its last column, which is the right-hand side of the order condition, when written in the form (IV.7.11’).





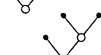
Convergence

Before we proceed to the actual construction of a new Rosenbrock method, we still have to study its convergence property. The following result will also involve

$$R(\infty) = 1 - b^T B^{-1} \mathbb{1} = 1 - \sum_{i,j} b_i \omega_{ij} \quad (4.25)$$

where $R(z)$ is the stability function (IV.7.14).

Table 4.1. Trees and elementary differentials

$\varrho(t)$	t	graph	$\gamma(t)$	$\Phi_j(t)$	$p_t(\gamma)$
4	t_{45}		4	$\sum \alpha_{jk} \alpha_{j\ell} \omega_{\ell m} \alpha_{mn} \alpha_{mp}$	1/4
2	u_{21}		1	$\sum \omega_{jk} \alpha_{k\ell} \alpha_{km}$	1
3	u_{31}		1	$\sum \omega_{jk} \alpha_{k\ell} \alpha_{km} \alpha_{kn}$	1
3	u_{32}		2	$\sum \omega_{jk} \alpha_{k\ell} \alpha_{km} \beta_{mn}$	$1/2 - \gamma$
3	u_{33}		1	$\sum \omega_{jk} \alpha_{k\ell} \alpha_{km} \omega_{mn} \alpha_{np} \alpha_{nq}$	1

We denote the local error of the Rosenbrock method (4.3) by

$$\delta y_h(x) = y_1 - y(x+h), \quad \delta z_h(x) = z_1 - z(x+h). \quad (4.26)$$

Here y_1, z_1 is the numerical solution obtained with the exact initial values $y_0 = y(x)$, $z_0 = z(x)$.

Theorem 4.10. *Suppose that g_z is regular in a neighbourhood of the solution $(y(x), z(x))$ of (4.1) and that the initial values (y_0, z_0) are consistent. If the stability function is such that $|R(\infty)| < 1$, and the local error satisfies*

$$\delta y_h(x) = \mathcal{O}(h^{p+1}), \quad \delta z_h(x) = \mathcal{O}(h^p), \quad (4.27)$$

then the Rosenbrock method (4.3) is convergent of order p ; i.e.,

$$y_n - y(x_n) = \mathcal{O}(h^p), \quad z_n - z(x_n) = \mathcal{O}(h^p) \quad \text{for} \quad x_n - x_0 = nh \leq \text{Const.}$$

Proof. Since g_z is regular we have

$$\|g_z^{-1}(y, z)g(y, z)\| \leq \delta \quad (4.28)$$

for (y, z) in a compact neighbourhood \mathcal{U} of the solution. The h -independent value of δ can be made arbitrarily small by shrinking \mathcal{U} . We also suppose for the moment that the numerical solution and all its internal stages remain in this neighbourhood. The propagation of local errors will be studied in part (a), and their accumulation over the whole interval in part (b).

a) We consider two pairs of initial values, (y_0, z_0) and $(\widehat{y}_0, \widehat{z}_0)$, and apply the method to each (these values may be inconsistent, but they are assumed to lie in \mathcal{U}). We shall prove that

$$\begin{aligned} \|y_1 - \widehat{y}_1\| &\leq (1 + hL)\|y_0 - \widehat{y}_0\| + hM\|z_0 - \widehat{z}_0\| \\ \|z_1 - \widehat{z}_1\| &\leq N\|y_0 - \widehat{y}_0\| + \kappa\|z_0 - \widehat{z}_0\| \end{aligned} \quad (4.29)$$

where $\kappa < 1$. For this we fix a sufficiently small step size h , and consider y_1, z_1 ,

k_i, ℓ_i as functions of (y_0, z_0) . We shall show that

$$\begin{aligned} \frac{\partial y_1}{\partial y_0} &= I + \mathcal{O}(h), & \frac{\partial y_1}{\partial z_0} &= \mathcal{O}(h), \\ \frac{\partial z_1}{\partial y_0} &= \mathcal{O}(1), & \frac{\partial z_1}{\partial z_0} &= R(\infty)I + \mathcal{O}(h + \delta). \end{aligned} \quad (4.30)$$

The mean value theorem then implies (4.29).

We first estimate k_i and ℓ_i , defined in (4.3b). Using (4.28) we compute ℓ_i from the second line and insert it into the first one. This yields successively $k_i = \mathcal{O}(h)$ and $\ell_i = \mathcal{O}(h + \delta)$ for all internal stages. We then differentiate (4.3b) once with respect to y_0 and once with respect to z_0 . An analysis similar to that for k_i and ℓ_i yields

$$\begin{aligned} \frac{\partial k_i}{\partial y_0} &= \mathcal{O}(h), & \frac{\partial k_i}{\partial z_0} &= \mathcal{O}(h) \\ \frac{\partial \ell_i}{\partial y_0} &= \mathcal{O}(1), & \frac{\partial \ell_i}{\partial z_0} &= - \sum_j \omega_{ij} I + \mathcal{O}(h + \delta) \end{aligned} \quad (4.31)$$

and the estimates (4.30) follow from (4.3a) and (4.25).

b) As a consequence of Lemma 3.9 (see Exercise 8), the propagation of the local errors $\delta y_h(x_{j-1}), \delta z_h(x_{j-1})$ to the solution at x_n can be bounded by

$$C(\|\delta y_h(x_{j-1})\| + (h + \kappa^{n-j})\|\delta z_h(x_{j-1})\|). \quad (4.32)$$

Summing up these terms from $j = 1$ to $j = n$ and using (4.27) gives the stated bounds for the global error, because $\sum_{j=1}^n (h + \kappa^{n-j}) \leq \text{Const.}$

Our assumption that the numerical solution and the internal stages lie in \mathcal{U} can now easily be justified by induction on the step number. The numerical solution remains $\mathcal{O}(h^p)$ -close to the exact solution and thus remains in \mathcal{U} for sufficiently small h . This implies $g(y_j, z_j) = \mathcal{O}(h^p)$ for all j and hence also $\ell_i = \mathcal{O}(h)$. Consequently (v_i, w_i) are also as close to the exact solution as we want. \square

Stiffly Accurate Rosenbrock Methods

We have already had several occasions to admire the beneficial effect of stiffly accurate Runge-Kutta methods (methods with $a_{si} = b_i$ for all i ; see Theorem 1.1 and Corollary 3.10). What is the corresponding condition for Rosenbrock methods?

Definition 4.11. A Rosenbrock method is called *stiffly accurate*, if

$$\alpha_{si} + \gamma_{si} = b_i \quad (i = 1, \dots, s) \quad \text{and} \quad \alpha_s = 1. \quad (4.33)$$

Recall that $\alpha_i = \sum_j \alpha_{ij}$. It has already been remarked at the end of Sect. IV.15 that methods satisfying (4.33) yield asymptotically exact results for the problem

$y' = \lambda(y - \varphi(x)) + \varphi'(x)$. A further interesting interpretation of this condition has been given by C. Schneider (1991). He argues that DAE's are combinations of differential equations and algebraic equations; hence methods should be equally valuable for both extreme cases, either a purely differential equation, or a purely algebraic equation

$$x' = 1, \quad 0 = g(x, z), \quad g_z \text{ invertible}. \quad (4.34)$$

Proposition 4.12. *A stiffly accurate Rosenbrock method, applied to (4.34), yields*

$$z_1 = w_s - g_z^{-1}(x_0, z_0) \cdot g(x_0 + h, w_s).$$

The numerical solution z_1 is thus the result of one simplified Newton iteration for $0 = g(x_0 + h, z)$ (with starting value w_s).

Proof. Condition (4.33) together with $\sum_i b_i = 1$ implies that $\gamma_s = \sum_j \gamma_{sj} = 0$. Therefore, the second line of (4.5) gives (observe that $k_i = h$ for the problem (4.34))

$$0 = g(x_0 + h, w_s) + g_z(x_0, z_0) \sum_{j=1}^i \gamma_{ij} \ell_j.$$

Inserting the expression thus obtained for $\sum_j \gamma_{ij} \ell_j$ into

$$z_1 = z_0 + \sum_{j=1}^s b_j \ell_j = w_s + \sum_{j=1}^s \gamma_{sj} \ell_j$$

proves the statement. \square

The values (v_s, w_s) of the last stage are often used as an embedded solution for step size control. If this is the case for a stiffly accurate method, then many of the algebraic order condition are automatically satisfied. This is a consequence of the following result.

Proposition 4.13. *Consider a stiffly accurate Rosenbrock method. For sufficiently regular problems (4.1) we have*

$$z_1 - z(x_0 + h) = \mathcal{O}(h^{q+1}) \quad (4.35)$$

if and only if

$$v_s - y(x_0 + h) = \mathcal{O}(h^q) \quad \text{and} \quad w_s - z(x_0 + h) = \mathcal{O}(h^q). \quad (4.36)$$

Proof. We use the characterization of Theorem 4.8 and the fact that (with ω_{ij} defined in (4.17))

$$\sum_i b_i \omega_{ij} = \begin{cases} 1 & \text{if } j = s \\ 0 & \text{else} \end{cases}. \quad (4.37)$$

Suppose first that (4.35) holds. For a tree $u = [\tau_y, t_2]_z$ with arbitrary $t_2 \in DAT_y$ we have, by definition of $\Phi_j(u)$ and $\gamma(u)$,

$$\sum_i b_i \Phi_i(u) = \sum_{i,j,k} b_i \omega_{ij} \alpha_j \alpha_{jk} \Phi_k(t_2) = \sum_k \alpha_{sk} \Phi_k(t_2) \quad (4.38)$$

and $\gamma(u) = \gamma(t_2)$. Consequently, the order condition is satisfied for u iff it is satisfied for t_2 . Since $\varrho(t_2) = \varrho(u) - 1$, we see that $v_s - y(x_0 + h) = \mathcal{O}(h^q)$ is a consequence of (4.35). By considering $u = [\tau_y, u_1]_z$ with $u_1 \in DAT_z$ we deduce the second relation of (4.36). The “if” part is proved in a similar way. \square

Finally we remark that because of (4.25) and (4.37) the stability function of a stiffly accurate Rosenbrock method always satisfies $R(\infty) = 0$. This is a desirable property when solving stiff or differential algebraic equations.

Construction of RODAS, a Stiffly Accurate Embedded Method

We want to construct an embedded Rosenbrock method (where $\hat{y}_1 = v_s$, $\hat{z}_1 = w_s$), such that both methods are stiffly accurate. This gives the following conditions

$$\begin{aligned} b_i &= \beta_{si} & (i = 1, \dots, s), & & \alpha_s &= 1 \\ \hat{b}_i &= \alpha_{si} = \beta_{s-1,i} & (i = 1, \dots, s-1), & & \alpha_{s-1} &= 1 \end{aligned} \quad (4.39)$$

(as usual $\beta_{ij} = \alpha_{ij} + \gamma_{ij}$). It follows from Proposition 4.12 that the last *two* stages represent simplified Newton iterations. Further, both methods have a stability function which vanishes at infinity. The construction of such a method of order 4(3) seems to be impossible with $s = 5$. We therefore put $s = 6$.

Here is the list of order conditions which have to be solved. We use the abbreviations α_i, β'_i defined in (IV.7.16), and the coefficients ω_{ij} from (4.17). We shall require that

$$y_1 - y(x_0 + h) = \mathcal{O}(h^5), \quad \hat{y}_1 - y(x_0 + h) = \mathcal{O}(h^4). \quad (4.40)$$

Since we have sufficiently many parameters we also require

$$v_{s-1} - y(x_0 + h) = \mathcal{O}(h^3), \quad w_{s-1} - z(x_0 + h) = \mathcal{O}(h^3). \quad (4.41)$$

By Proposition 4.13 this implies

$$\hat{z}_1 - z(x_0 + h) = \mathcal{O}(h^4), \quad z_1 - z(x_0 + h) = \mathcal{O}(h^5), \quad (4.42)$$

which is more than sufficient to ensure convergence of order 4 (see Theorem 4.10). The conditions for (4.40) and (4.41) are (see Table IV.7.1 and Table 4.1)

$$b_1 + b_2 + b_3 + b_4 + (b_5 + b_6) = 1 \quad (4.43a)$$

$$b_2 \beta'_2 + b_3 \beta'_3 + b_4 \beta'_4 + (b_5 + b_6)(1 - \gamma) = \frac{1}{2} - \gamma \quad (4.43b)$$

$$b_2 \alpha_2^2 + b_3 \alpha_3^2 + b_4 \alpha_4^2 + (b_5 + b_6) = \frac{1}{3} \quad (4.43c)$$

$$b_3\beta_{32}\beta'_2 + b_4 \sum' \beta_{4i}\beta'_i + (b_5 + b_6)\left(\frac{1}{2} - 2\gamma + \gamma^2\right) = \frac{1}{6} - \gamma + \gamma^2 \quad (4.43d)$$

$$b_2\alpha_2^3 + b_3\alpha_3^3 + b_4\alpha_4^3 + (b_5 + b_6) = \frac{1}{4} \quad (4.43e)$$

$$b_3\alpha_3\alpha_{32}\beta'_2 + b_4\alpha_4 \sum' \alpha_{4i}\beta'_i + (b_5 + b_6)\left(\frac{1}{2} - \gamma\right) = \frac{1}{8} - \frac{\gamma}{3} \quad (4.43f)$$

$$b_3\beta_{32}\alpha_2^2 + b_4 \sum' \beta_{4i}\alpha_i^2 + (b_5 + b_6)\left(\frac{1}{3} - \gamma\right) = \frac{1}{12} - \frac{\gamma}{3} \quad (4.43g)$$

$$b_4\beta_{43}\beta_{32}\beta'_2 + (b_5 + b_6)\left(\frac{1}{6} - \frac{3}{2}\gamma + 3\gamma^2 - \gamma^3\right) = \frac{1}{24} - \frac{\gamma}{2} + \frac{3}{2}\gamma^2 - \gamma^3 \quad (4.43h)$$

$$b_3\alpha_3\alpha_{32}\omega_{22}\alpha_2^2 + b_4\alpha_4 \sum_{i,j} \alpha_{4i}\omega_{ij}\alpha_j^2 + (b_5 + b_6) = \frac{1}{4} \quad (4.43i)$$

$$\alpha_{62}\beta'_2 + \alpha_{63}\beta'_3 + \alpha_{64}\beta'_4 = \frac{1}{2} - 2\gamma + \gamma^2 \quad (4.43j)$$

$$\alpha_{62}\alpha_2^2 + \alpha_{63}\alpha_3^2 + \alpha_{64}\alpha_4^2 = \frac{1}{3} - \gamma \quad (4.43k)$$

$$\alpha_{63}\beta_{32}\beta'_2 + \alpha_{64} \sum' \beta_{4i}\beta'_i = \frac{1}{6} - \frac{3}{2}\gamma + 3\gamma^2 - \gamma^3 \quad (4.43l)$$

$$\alpha_{52}\beta'_2 + \alpha_{53}\beta'_3 + \alpha_{54}\beta'_4 = \frac{1}{2} - \gamma \quad (4.43m)$$

$$\sum_{i=1}^4 \alpha_{5i} \sum_{j=1}^i \omega_{ij} \alpha_j^2 = 1 \quad (4.43n)$$

In order to solve the system (4.39), (4.43a–n) we can take γ , α_2 , α_3 , α_4 , $\beta'_2 = \beta_{21}$, β'_3 , β'_4 as free parameters. The remaining coefficients can then be computed as follows:

Step 1. We have $b_6 = \gamma$ by (4.39). The remaining b_i can be chosen such that (4.43a,b,c,e) are satisfied. We have one degree of freedom which can be exploited to fulfill the additional order condition $\sum_i b_i \alpha_i^4 = 1/5$. This step also yields $\beta_{6i} = b_i$ for $i = 1, \dots, 6$.

Step 2. Compute the two expressions $b_3\beta_{32} + b_4\beta_{42}$ and $b_4\beta_{43}$ from (4.43d,g), and then β_{32} from (4.43h). Because of $\beta'_i = \sum_{j=1}^{i-1} \beta_{ij}$ this determines all β_{ij} with $i \leq 4$. Observe that $\beta_{ii} = \gamma$ for all i .

Step 3. Solve the linear system (4.43j,k,l) for α_{62} , α_{63} , α_{64} . We have $\alpha_{65} = \gamma$ by (4.39) and compute α_{61} from $\alpha_6 = \sum_i \alpha_{6i} = 1$. This also yields $\beta_{5i} = \alpha_{6i}$ by (4.39). Hence all β_{ij} and ω_{ij} , and also $\hat{b}_i = \beta_{5i}$ ($i = 1, \dots, 5$) are determined at this stage.

Step 4. The conditions (4.43m,n) and $\alpha_5 = 1$ constitute 3 linear equations in the four unknown parameters α_{51} , α_{52} , α_{53} , α_{54} . We have one degree of freedom in this step.

Step 5. The remaining two conditions (4.43f,i) are linear equations in α_{32} , α_{42} , α_{43} . We have one more degree of freedom which can be exploited to fulfill the order condition for the tree $[\tau_y, \tau_y, [\tau_y]_y]_y$. The values of α_{i1} are then determined by $\alpha_i = \sum_{j=1}^{i-1} \alpha_{ij}$, and those of γ_{ij} are given by $\gamma_{ij} = \beta_{ij} - \alpha_{ij}$.

The coefficients for the code RODAS of the appendix were computed with the above procedure. In step 4 we have added the condition

$$\sum_{i,j} \alpha_{5i} \omega_{ij} = 1 \quad (4.44)$$

which will be explained in Exercise 3 below. The free parameters were chosen in

order to get an A -stable method with small error constants. The result is

$$\begin{aligned}\gamma &= 0.25 \\ \alpha_2 &= 0.386 & \alpha_3 &= 0.21 & \alpha_4 &= 0.63 \\ \beta'_2 &= 0.0317 & \beta'_3 &= 0.0635 & \beta'_4 &= 0.3438\end{aligned}\quad (4.45)$$

We do not claim that these values are optimal. Nevertheless, the numerical results of Sect. IV.10 (Fig. IV.10.8, IV.10.9 and IV.10.12) are encouraging. Although the new method needs 6 function evaluations per step, it is in general superior to the classical methods of Table IV.7.2 which need only 3 evaluations per step.

A different set of coefficients, based on the same construction, has been proposed by Steinebach (1995). The free parameters are chosen in order to satisfy the Scholz conditions $C_2(z) \equiv 0$ and $C_3(z) \equiv 0$ (see Eq. (15.41) of Sect. IV.15).

Dense Output. A natural way to define a continuous numerical solution for $y(x_0 + \theta h)$, $z(x_0 + \theta h)$ is

$$y_1(\theta) = y_0 + \sum_{i=1}^s b_i(\theta)k_i, \quad z_1(\theta) = z_0 + \sum_{i=1}^s b_i(\theta)\ell_i, \quad (4.46)$$

where the $b_i(\theta)$ are polynomials which satisfy $b_i(0) = 0$, $b_i(1) = b_i$. In complete analogy to Theorem 4.8 we have

$$\begin{aligned}y(x_0 + \theta h) - y_1(\theta) &= \mathcal{O}(h^{p+1}) \quad \text{iff} \quad \sum_{i=1}^s b_i(\theta)\Phi_i(t) = \frac{\theta \varrho(t)}{\gamma(t)} \\ &\quad \text{for } t \in DAT_y, \varrho(t) \leq p, \\ z(x_0 + \theta h) - z_1(\theta) &= \mathcal{O}(h^{q+1}) \quad \text{iff} \quad \sum_{i=1}^s b_i(\theta)\Phi_i(u) = \frac{\theta \varrho(u)}{\gamma(u)} \\ &\quad \text{for } u \in DAT_z, \varrho(u) \leq q.\end{aligned}\quad (4.47)$$

In our situation ($s = 6$) it is easy to fulfill these conditions with $p = 3$ and $q = 2$. The additional condition $b_s(\theta) = \gamma\theta$ makes the solution unique.

Methods of Order 5. C. Schneider (1991b) first constructed stiffly accurate Rosenbrock methods of order 5 with $s = 8$ stages. Di Marzo (1992) then determined carefully the free parameters to obtain A -stability and small error constants. The resulting method, implemented in the code RODAS5, gives excellent results (see Sect. IV.10).

Inconsistent Initial Values

Even if we start the computation with consistent initial values, the numerical solution (y_n, z_n) of a Rosenbrock method does not, in general, satisfy $g(y_n, z_n) = 0$. It is therefore of interest to investigate the local error also for inconsistent initial

values (y_0, z_0) . But what is the local error? To which solution of (4.1) should we compare the numerical values? If

$$\|(g_z^{-1}g)(y_0, z_0)\| \leq \delta \quad (4.48)$$

with sufficiently small δ , we can find (because of (1.7)) a locally unique \widehat{z}_0 which satisfies $g(y_0, \widehat{z}_0) = 0$. It is natural to compare the numerical solution (y_1, z_1) to that solution of (4.1) which passes through (y_0, \widehat{z}_0) .

Our first aim is to write this solution in terms of elementary differentials evaluated at (y_0, z_0) . Using

$$\widehat{z}_0 - z_0 = (-g_z^{-1}g)(y_0, z_0) + \mathcal{O}(\delta^2),$$

which is a consequence of $0 = g(y_0, z_0) + g_z(y_0, z_0)(\widehat{z}_0 - z_0) + \dots$, we get

$$y(x_0 + h) = y_0 + hf(y_0, \widehat{z}_0) + \mathcal{O}(h^2) \quad (4.49)$$

$$= y_0 + hf(y_0, z_0) + h(f_z(-g_z^{-1}g)(y_0, z_0) + \mathcal{O}(h^2 + h\delta^2))$$

$$z(x_0 + h) = \widehat{z}_0 + h(-g_z^{-1}g_y f)(y_0, \widehat{z}_0) + \mathcal{O}(h^2) \quad (4.50)$$

$$\begin{aligned} &= z_0 + (-g_z^{-1}g)(y_0, z_0) + h(-g_z^{-1}g_y f)(y_0, z_0) \\ &\quad + h(-g_z^{-1}g_{zz}(-g_z^{-1}g, -g_z^{-1}g_y f))(y_0, z_0) \\ &\quad + h(-g_z^{-1}g_{yz}(f, -g_z^{-1}g))(y_0, z_0) \\ &\quad + h(-g_z^{-1}g_y f_z(-g_z^{-1}g)(y_0, z_0) + \mathcal{O}(h^2 + \delta^2)) \end{aligned}$$

The expressions so obtained allow a nice interpretation using trees. We only have to add in the recursive Definition 4.1 a tree of order 0, which consists of a fat root. We denote this tree by \emptyset_z , and extend Definition 4.3 by setting $F(\emptyset_z)(y, z) = (-g_z^{-1}g)(y, z)$. Then, the expressions of (4.49) and (4.50) correspond to the trees of Fig. 4.4.



Fig. 4.4. Trees, to be considered for inconsistent initial values

The numerical solution also possesses an expansion of the form (4.49), (4.50) with additional method-dependent coefficients. The first few terms are as follows:




$$y_1 = y_0 + \left(\sum_i b_i \right) hf(y_0, z_0) + \left(\sum_{i,j,k} b_i \beta_{ij} \omega_{jk} \right) h(f_z(-g_z^{-1}g)(y_0, z_0) + \mathcal{O}(h^2 + h\delta^2))$$

$$z_1 = z_0 + \left(\sum_{i,j} b_i \omega_{ij} \right) (-g_z^{-1}g)(y_0, z_0) + \mathcal{O}(h + \delta^2).$$

In order to understand the form of these new coefficients we have to extend the proof of Theorem 4.6. It turns out that the elementary differentials are multiplied by $\gamma(t) \sum_i b_i \Phi_i(t)$ or $\gamma(u) \sum_i b_i \Phi_i(u)$, where γ and Φ_i are defined by $\gamma(\emptyset_z) =$

1, $\Phi_i(\emptyset_z) = \sum_j \omega_{ij}$ and the recursion of Theorem 4.6. Equating the coefficients of the exact and numerical solutions yields new order conditions for the case of inconsistent initial values. The first of these (to be added to those of Table IV.7.1 and Table 4.1) are presented in Table 4.2.

Table 4.2. Order conditions for inconsistent initial values

tree	order condition	size of error term
	$\sum b_i \alpha_i \alpha_{ij} \omega_{jk} = 1/2$	$\mathcal{O}(h^2 \delta)$
	$\sum b_i \omega_{ij} = 1$	$\mathcal{O}(\delta)$
	$\sum b_i \omega_{ij} \alpha_j \alpha_{jk} \omega_{kl} = 1$	$\mathcal{O}(h \delta)$

Remarks. a) The first condition of Table 4.2 is exactly the same as that found by van Veldhuizen (1984) in a different context. It implies that the local error of the y -component is of size $\mathcal{O}(h^{p+1} + h^3 \delta + h \delta^2)$.

b) Condition $\sum_{i,j} b_i \omega_{ij} = 1$ means that the stability function satisfies $R(\infty) = 0$. Unless this condition is satisfied, the local error of the z -component contains an h -independent term of size δ (which usually is near to Tol). This was observed numerically in Fig. IV.7.4 and explains the phenomenon of Fig. IV.7.3.

c) For Rosenbrock methods which satisfy (4.39), the second and third conditions of Table 4.2 are automatically fulfilled. For such methods the local error of the z -component is of size $\mathcal{O}(h^{q+1} + h^2 \delta + \delta^2)$.

Exercises

1. (Roche 1989). Consider the implicit Runge-Kutta method (1.11) applied to (1.6).

a) Prove that $z_1 - z(x_0 + h) = \mathcal{O}(h^{q+1})$ iff

$$\sum_{i=1}^s b_i \Phi_i(u) = \frac{1}{\gamma(u)} \quad \text{for } u \in DAT_z, \varrho(u) \leq q,$$

where $\gamma(u)$ and $\Phi_i(u)$ are defined as in Theorem 4.6, but all coefficients α_{ij} and β_{ij} are replaced by the Runge-Kutta coefficients a_{ij} .

b) Show that those trees in DAT_z which have more than one fat vertex, are redundant.

2. The simplifying assumptions (4.39) imply that many of the (algebraic) order conditions are automatically satisfied. Characterize the corresponding trees.

3. State the order condition for the tree $[\tau_y, [\tau_y, \emptyset_z]_z]_z$.
- a) Show that the corresponding error term is of size $\mathcal{O}(h^2\delta)$ with δ given in (4.48).
- b) For methods satisfying (4.39), this condition is equivalent to (4.44).
4. (Ostermann 1990). Suppose that the Rosenbrock method (4.3) satisfies (4.27). Define polynomials $b_i(\theta)$ of degree $q = [(p+1)/2]$ by $b_i(0) = 0$, $b_i(1) = b_i$, and

$$\int_0^1 b_i(\theta) \theta^{\ell-1} d\theta = \begin{cases} \sum_j b_j(\alpha_{ji} + \gamma_{ji}) & \text{if } \ell = 1 \\ \sum_j b_j \alpha_j^{\ell-1} \alpha_{ji} & \text{if } \ell = 2, \dots, q-1. \end{cases}$$

Prove that the error of the dense output formulas (4.46) is $\mathcal{O}(h^{q+1})$.

Hint. Extend the ideas of Exercise II.17.5 to Rosenbrock methods.

5. Suppose that a Rosenbrock method is implemented in the form (IV.7.25). If it satisfies (4.39), then its last two stages allow a very simple implementation

Hint. Prove that

$$m_i = \begin{cases} a_{si} & i = 1, \dots, s-1 \\ 1 & i = s, \end{cases} \quad a_{si} = \begin{cases} a_{s-1,i} & i = 1, \dots, s-2 \\ 1 & i = s-1. \end{cases}$$

6. *Partitioned Rosenbrock methods* (Rentrop, Roche & Steinebach 1989). Consider the method (4.3) with f_y and f_z replaced by 0. Derive necessary and sufficient conditions that it be of order p .

Remark. Case (a) of Lemma 4.9 remains valid in this new situation. However, the trees of Lemma 4.9b give rise to new conditions.

7. What is the “algebraic order” of the classical 4th order Rosenbrock methods of Section IV.7?
8. Let $\{u_n\}$, $\{v_n\}$ be two sequences of non-negative numbers satisfying (componentwise)

$$\begin{pmatrix} u_{n+1} \\ v_{n+1} \end{pmatrix} \leq \begin{pmatrix} 1+hL & hM \\ N & \kappa \end{pmatrix} \begin{pmatrix} u_n \\ v_n \end{pmatrix}$$

with $0 \leq \kappa < 1$ and positive constants L, M, N . Prove that for $h \leq h_0$ and $nh \leq \text{Const}$

$$u_n \leq C(u_0 + hv_0), \quad v_n \leq C(u_0 + (h + \kappa^n)v_0).$$

Hint. Apply Lemma 3.9 with $\varepsilon = h$ and $M = 0$.

VI.5 Extrapolation Methods

The numerical computations of Sect. IV.10 have revealed the extrapolation code SEULEX as one of the best method for very stringent tolerances. The aim of the present section is to justify theoretically the underlying numerical method, the extrapolated linearly implicit Euler method, for singular perturbation problems as a representative of stiff equations.

Linearly Implicit Euler Discretization

The linearly implicit Euler method (IV.9.25) applied to the singular perturbation problem (1.5) reads

$$\begin{pmatrix} I - hf_y(0) & -hf_z(0) \\ -hg_y(0) & \varepsilon I - hg_z(0) \end{pmatrix} \begin{pmatrix} y_{i+1} - y_i \\ z_{i+1} - z_i \end{pmatrix} = h \begin{pmatrix} f(y_i, z_i) \\ g(y_i, z_i) \end{pmatrix}. \quad (5.1)$$

Here we have used abbreviations such as $f_y(0) = f_y(y_0, z_0)$ for the partial derivatives. We recall that the numerical approximations at $x_0 + H$ ($H = nh$) are extrapolated according to (IV.9.26).

For the differential algebraic problem (1.6) we just put $\varepsilon = 0$ in (5.1). This yields

$$\begin{pmatrix} I - hf_y(0) & -hf_z(0) \\ -hg_y(0) & -hg_z(0) \end{pmatrix} \begin{pmatrix} y_{i+1} - y_i \\ z_{i+1} - z_i \end{pmatrix} = h \begin{pmatrix} f(y_i, z_i) \\ g(y_i, z_i) \end{pmatrix}. \quad (5.2)$$

Possible extensions to non-autonomous problems have been presented in Sect. IV.9. For problems $Mu' = \varphi(u)$ we use the formulation (IV.9.34) also for singular M . Due to the invariance of the method with respect to the transformation (1.23), all results of this section are equally valid for $Mu' = \varphi(u)$ of index 1.

The performance of extrapolation methods relies heavily on the existence of an asymptotic expansion of the global error. Such expansions are well understood, if the differential equation is nonstiff (see Sections II.8 and IV.9). But what happens if the problem is stiff or differential-algebraic?

Continued study of special problems is still a commendable way
towards greater insight . . . (E. Hopf 1950)

Example 5.1. Consider the test problem

$$y' = 1, \quad \varepsilon z' = -z + g(y). \quad (5.3)$$

Method (5.1) yields the exact result $y_i = x_i = x_0 + ih$ for the y -component, and the recursion

$$(\varepsilon + h)z_{i+1} = \varepsilon z_i + hg(x_i) + h^2 g'(x_0) \quad (5.4)$$

for the z -component. In order to compute the coefficients of the asymptotic expansion (Theorem II.8.1), we insert

$$z_i = z(x_i) + hb_1(x_i) + h^2 b_2(x_i) + h^3 b_3(x_i) + \dots \quad (5.5)$$

into (5.4), expand into a Taylor series and compare the coefficients of h^j . This yields the differential equation

$$\varepsilon b_1'(x) + b_1(x) = -\frac{\varepsilon}{2} z''(x) - z'(x) + g'(x_0)$$

for $b_1(x)$, and similar ones for $b_2(x)$, $b_3(x)$, etc. Putting $i = 0$ in (5.5) we get the initial values $b_i(x_0) = 0$ (all i). In general, the computation of the functions $b_1(x)$, $b_2(x)$, \dots is rather tedious. We therefore continue this example for the special case $x_0 = 0$, $g(x) = x^2 + 2\varepsilon x$, and $z_0 = 0$, so that the exact solution of (5.3) is $z(x) = x^2$. In this situation we get

$$\begin{aligned} b_1(x) &= -3\varepsilon e^{-x/\varepsilon} + 3\varepsilon - 2x \\ b_2(x) &= -\left(1 + \frac{3x}{2\varepsilon}\right) e^{-x/\varepsilon} + 1 \\ b_3(x) &= \left(\frac{x}{2\varepsilon^2} - \frac{3x^2}{8\varepsilon^3}\right) e^{-x/\varepsilon} \end{aligned} \quad (5.6)$$

etc. We observe that for $\varepsilon \rightarrow 0$, the function $b_2(x)$ becomes discontinuous at $x = 0$, and $b_3(x)$ is even not uniformly bounded. Hence, the expansion (5.5) is not useful for the study of extrapolation, if ε is small compared to the step size H .

The idea is now to omit in (5.6) the terms containing the factor $e^{-x/\varepsilon}$ by requiring that the functions $b_i(x)$ be *smooth* uniformly in ε and, instead, to add a discrete perturbation β_i to (5.5). For our example, this then becomes

$$z_i = x_i^2 + h(3\varepsilon - 2x_i) + h^2 + \beta_i. \quad (5.7a)$$

Inserting (5.7a) into (5.4) gives the relation $(\varepsilon + h)\beta_{i+1} = \varepsilon\beta_i$. The value of β_0 is obtained from (5.7a) with $i = 0$. We thus get

$$\beta_i = -\left(1 + \frac{h}{\varepsilon}\right)^{-i} (3\varepsilon h + h^2). \quad (5.7b)$$

If the numerical solution is extrapolated, the smooth terms in (5.7) are eliminated one after the other. It remains to study the effect of extrapolation on the perturbation terms β_i . If the differential equation is very stiff ($\varepsilon \ll h$), these terms are very small and may be neglected over a wide range of h (observe that $i \geq n_1$).

Example 5.2. For the differential-algebraic problem

$$y' = 1, \quad 0 = -z + g(y) \quad (5.8)$$

with initial values $y_0 = x_0$, $z_0 = g(x_0)$ the numerical solution, given by (5.2), is

$$z_i = \begin{cases} g(x_0) & \text{for } i = 0 \\ g(x_{i-1}) + hg'(x_0) & \text{for } i \geq 1. \end{cases}$$

Developing its second formula (for $i \geq 1$) yields

$$z_i = g(x_i) + h(g'(x_0) - g'(x_i)) + \frac{h^2}{2}g''(x_i) - \frac{h^3}{6}g'''(x_i) + \mathcal{O}(h^4).$$

If we add the perturbation

$$\beta_i = h\beta_i^1 + h^2\beta_i^2 + h^3\beta_i^3 + \dots \quad (5.9)$$

(which is different from zero only for $i = 0$) we get for *all* i

$$z_i - g(x_i) = \sum_{j=1}^3 h^j(b_j(x_i) + \beta_i^j) + \mathcal{O}(h^4) \quad (5.10)$$

where

$$b_1(x) = g'(x_0) - g'(x), \quad b_2(x) = \frac{1}{2}g''(x), \quad b_3(x) = -\frac{1}{6}g'''(x)$$

are smooth functions and the perturbations are given by

$$\beta_0^1 = 0, \quad \beta_0^2 = -\frac{1}{2}g''(x_0), \quad \beta_0^3 = \frac{1}{6}g'''(x_0).$$

If we add a further algebraic equation to (5.8), e.g., $0 = u - k(z)$, and again apply Method (5.2), we get three different formulas for u_i , one for $i = 0$, one for $i = 1$, and a different one for $i \geq 2$. In an expansion of the type (5.10) for $u_i - k(g(x_i))$, perturbation terms will be present for $i = 0$ and for $i = 1$.

Perturbed Asymptotic Expansion

For general differential algebraic problems we have the following result.

Theorem 5.3 (Deuffhard, Hairer & Zugck 1987). *Consider the problem (1.6) with consistent initial values (y_0, z_0) , and suppose that (1.7) is satisfied. The global error of the linearly implicit Euler method (5.2) then has an asymptotic h -expansion of the form*

$$\begin{aligned} y_i - y(x_i) &= \sum_{j=1}^M h^j (a_j(x_i) + \alpha_i^j) + \mathcal{O}(h^{M+1}) \\ z_i - z(x_i) &= \sum_{j=1}^M h^j (b_j(x_i) + \beta_i^j) + \mathcal{O}(h^{M+1}) \end{aligned} \quad (5.11)$$

where $a_j(x)$, $b_j(x)$ are smooth functions and the perturbations satisfy (see Table 5.1 and 5.2)

$$\alpha_i^1 = 0, \quad \alpha_i^2 = 0, \quad \alpha_i^3 = 0, \quad \beta_i^1 = 0 \quad \text{for } i \geq 0 \quad (5.12a)$$

$$\beta_i^2 = 0 \quad \text{for } i \geq 1 \quad (5.12b)$$

$$\alpha_i^j = 0 \quad \text{for } i \geq j-4 \quad \text{and } j \geq 4 \quad (5.12c)$$

$$\beta_i^j = 0 \quad \text{for } i \geq j-2 \quad \text{and } j \geq 3. \quad (5.12d)$$

The error terms in (5.11) are uniformly bounded for $x_i = ih \leq H$, if H is sufficiently small.

Table 5.1. Non-zero α 's

	h	h^2	h^3	h^4	h^5	h^6	h^7
y_0	0	0	0	0	*	*	*
y_1	0	0	0	0	0	*	*
y_2	0	0	0	0	0	0	*
y_3	0	0	0	0	0	0	0
y_4	0	0	0	0	0	0	0
y_5	0	0	0	0	0	0	0

Table 5.2. Non-zero β 's

	h	h^2	h^3	h^4	h^5	h^6	h^7
z_0	0	*	*	*	*	*	*
z_1	0	0	0	*	*	*	*
z_2	0	0	0	0	*	*	*
z_3	0	0	0	0	0	*	*
z_4	0	0	0	0	0	0	*
z_5	0	0	0	0	0	0	0

Proof. In part (a) we shall recursively construct truncated expansions

$$\begin{aligned} \widehat{y}_i &= y(x_i) + \sum_{j=1}^M h^j (a_j(x_i) + \alpha_i^j) + h^{M+1} \alpha_i^{M+1} \\ \widehat{z}_i &= z(x_i) + \sum_{j=1}^M h^j (b_j(x_i) + \beta_i^j) \end{aligned} \quad (5.13)$$

such that the defect of \widehat{y}_i , \widehat{z}_i inserted into the method is small; more precisely, we require that

$$\begin{pmatrix} I - hf_y(0) & -hf_z(0) \\ -hg_y(0) & -hg_z(0) \end{pmatrix} \begin{pmatrix} \widehat{y}_{i+1} - \widehat{y}_i \\ \widehat{z}_{i+1} - \widehat{z}_i \end{pmatrix} = h \begin{pmatrix} f(\widehat{y}_i, \widehat{z}_i) \\ g(\widehat{y}_i, \widehat{z}_i) \end{pmatrix} + O(h^{M+2}). \quad (5.14)$$

For the initial values we require $\widehat{y}_0 = y_0$, $\widehat{z}_0 = z_0$, or equivalently

$$a_j(0) + \alpha_0^j = 0, \quad b_j(0) + \beta_0^j = 0, \quad (5.15)$$

and the perturbation terms are assumed to satisfy

$$\alpha_i^j \rightarrow 0, \quad \beta_i^j \rightarrow 0 \quad \text{for } i \rightarrow \infty, \quad (5.16)$$

otherwise, these limits could be added to the smooth parts. The result will then follow from a stability estimate derived in part (b).

a) For the construction of $a_j(x)$, $b_j(x)$, α_i^j , β_i^j we insert (5.13) into (5.14), and develop

$$\begin{aligned} f(\widehat{y}_i, \widehat{z}_i) &= f(y(x_i), z(x_i)) + f_y(x_i)(ha_1(x_i) + h\alpha_i^1 + \dots) \\ &\quad + f_z(x_i)(hb_1(x_i) + h\beta_i^1 + \dots) \\ &\quad + f_{yy}(x_i)(ha_1(x_i) + h\alpha_i^1 + \dots)^2 + \dots, \end{aligned}$$

$$\begin{aligned} \widehat{y}_{i+1} - \widehat{y}_i &= y(x_{i+1}) - y(x_i) + h(a_1(x_{i+1}) - a_1(x_i) + \alpha_{i+1}^1 - \alpha_i^1) + \dots \\ &= hy'(x_i) + \frac{h^2}{2}y''(x_i) + \dots + h^2a_1'(x_i) + h(\alpha_{i+1}^1 - \alpha_i^1) + \dots, \end{aligned}$$

where $f_y(x) = f_y(y(x), z(x))$, etc. Similarly, we develop $g(\widehat{y}_i, \widehat{z}_i)$ and $\widehat{z}_{i+1} - \widehat{z}_i$, and compare coefficients of h^{j+1} (for $j = 0, \dots, M$). Each power of h will lead to *two* conditions — one containing the smooth functions and the other containing the perturbation terms.

First step. Equating the coefficients of h^1 yields the equations (1.6) for the smooth part (due to consistency of the method), and $\alpha_{i+1}^1 - \alpha_i^1 = 0$ for $i \geq 0$. Because of (5.16) we get $\alpha_i^1 = 0$ for all $i \geq 0$ (compare (5.12a)).

Second step. The coefficients of h^2 give

$$a_1'(x) + \frac{1}{2}y''(x) - f_y(0)y'(x) - f_z(0)z'(x) = f_y(x)a_1(x) + f_z(x)b_1(x) \quad (5.17a)$$

$$-g_y(0)y'(x) - g_z(0)z'(x) = g_y(x)a_1(x) + g_z(x)b_1(x) \quad (5.17b)$$

$$\alpha_{i+1}^2 - \alpha_i^2 - f_z(0)(\beta_{i+1}^1 - \beta_i^1) = f_z(0)\beta_i^1 \quad (5.17c)$$

$$-g_z(0)(\beta_{i+1}^1 - \beta_i^1) = g_z(0)\beta_i^1. \quad (5.17d)$$

Observe that the coefficients α_i^ℓ , β_i^ℓ have to be independent of h , so that $f_z(0)$, $g_z(0)$ cannot be replaced by $f_z(x_i)$, $g_z(x_i)$ in the right-hand sides of (5.17c, d). The system (5.17) can be solved as follows. Compute $b_1(x)$ from (5.17b) and insert it into (5.17a). This gives a linear differential equation for $a_1(x)$. Because of (5.15) and $\alpha_0^1 = 0$ the initial value is $a_1(0) = 0$. Therefore $a_1(x)$ and $b_1(x)$ are uniquely determined by (5.17a, b). Differentiating $g(y(x), z(x)) = 0$ and putting $x = 0$ implies that the left-hand side of (5.17b) vanishes at $x = 0$. Consequently, we have $b_1(0) = 0$ and by (5.15), also $\beta_0^1 = 0$. Condition (5.17d) then implies $\beta_i^1 = 0$ (all i), and (5.17c) together with (5.16) give $\alpha_i^2 = 0$ (all i).

Third step. As in the second step we get (for $j = 2$)

$$a_j'(x) = f_y(x)a_j(x) + f_z(x)b_j(x) + r(x) \quad (5.18a)$$

$$0 = g_y(x)a_j(x) + g_z(x)b_j(x) + s(x), \quad (5.18b)$$

where $r(x)$, $s(x)$ are known functions depending on derivatives of $y(x)$, $z(x)$, and on $a_\ell(x)$, $b_\ell(x)$ with $\ell \leq j - 1$. We further get

$$\alpha_{i+1}^3 - \alpha_i^3 = f_z(0)\beta_{i+1}^2 \quad (5.18c)$$

$$0 = g_z(0)\beta_{i+1}^2. \quad (5.18d)$$

We compute $a_2(x)$, $b_2(x)$ as in step 2. However, $b_2(0) \neq 0$ in general, and for the first time, we are forced to introduce a perturbation term $\beta_0^2 \neq 0$. From (5.18c, d) we then get $\beta_i^2 = 0$ (for $i \geq 1$) and $\alpha_i^3 = 0$ (for all i).

Fourth step. Comparing the coefficients of h^4 we just get (5.18a,b) with $j = 3$ and (5.18c,d) with the upper index raised by 1. As above we conclude $\beta_i^3 = 0$ (for $i \geq 1$) and $\alpha_i^4 = 0$ (for all i).

General step. The conditions for the smooth functions are (5.18a,b). For the perturbation terms we get

$$\alpha_{i+1}^{j+1} - \alpha_i^{j+1} = f_z(0)\beta_{i+1}^j + \varrho_i^j \quad (5.19c)$$

$$0 = g_z(0)\beta_{i+1}^j + \sigma_i^j, \quad (5.19d)$$

where ϱ_i^j , σ_i^j are linear combinations of expressions which contain as factors α_{i+1}^ℓ , $\alpha_i^{\ell-1}$, $\beta_i^{\ell-1}$ with $\ell \leq j$. For example, we have $\varrho_i^4 = f_{zz}(0)(\beta_i^2)^2$ and $\sigma_i^4 = g_{zz}(0)(\beta_i^2)^2$. The proof of (5.12) is now by induction on j . By the induction hypothesis we have $\varrho_i^j = 0$, $\sigma_i^j = 0$ for $i \geq j - 3$. Formula (5.19d) hence implies $\beta_{i+1}^j = 0$ (for $i \geq j - 3$) and (5.19c) together with (5.16) gives $\alpha_i^{j+1} = 0$ (for $i \geq j - 3$). But this is simply the statement (5.12c,d).

b) We still have to estimate the remainder term, i.e., differences $\Delta y_i = y_i - \hat{y}_i$, $\Delta z_i = z_i - \hat{z}_i$. Subtracting (5.14) from (5.2) and eliminating Δy_{i+1} , Δz_{i+1} yields

$$\begin{pmatrix} \Delta y_{i+1} \\ \Delta z_{i+1} \end{pmatrix} = \begin{pmatrix} \Delta y_i \\ \Delta z_i \end{pmatrix} + \begin{pmatrix} I + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & -g_z^{-1}(0) \end{pmatrix} \begin{pmatrix} h(f(y_i, z_i) - f(\hat{y}_i, \hat{z}_i)) \\ g(y_i, z_i) - g(\hat{y}_i, \hat{z}_i) \end{pmatrix} + \begin{pmatrix} \mathcal{O}(h^{M+2}) \\ \mathcal{O}(h^{M+1}) \end{pmatrix}.$$

The application of a Lipschitz condition for $f(y, z)$ and $g(y, z)$ then gives

$$\begin{pmatrix} \|\Delta y_{i+1}\| \\ \|\Delta z_{i+1}\| \end{pmatrix} \leq \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & \varrho \end{pmatrix} \begin{pmatrix} \|\Delta y_i\| \\ \|\Delta z_i\| \end{pmatrix} + \begin{pmatrix} \mathcal{O}(h^{M+2}) \\ \mathcal{O}(h^{M+1}) \end{pmatrix}, \quad (5.20)$$

where $|\varrho| < 1$ if H is sufficiently small. Applying Lemma 3.9 we deduce $\|\Delta y_i\| + \|\Delta z_i\| = \mathcal{O}(h^{M+1})$. \square

Order Tableau

We consider (5.2) as our basic method for extrapolation, i.e., we take some step number sequence $n_1 < n_2 < \dots$, put $h_j = H/n_j$, and define

$$Y_{j1} = y_{h_j}(x_0 + H), \quad Z_{j1} = z_{h_j}(x_0 + H), \quad (5.21)$$

the numerical solution of (1.6) after n_j steps with step size h_j . We then extrapolate these values according to (IV.9.26) and obtain Y_{jk} , Z_{jk} . What is the order of the approximations thus obtained?

Theorem 5.4 (Deufhard, Hairer & Zugck 1987). *If we consider the harmonic sequence $\{1, 2, 3, 4, \dots\}$, then the extrapolated values Y_{jk} , Z_{jk} satisfy*

$$Y_{jk} - y(x_0 + h) = \mathcal{O}(H^{r_{jk}+1}), \quad Z_{jk} - z(x_0 + H) = \mathcal{O}(H^{s_{jk}}) \quad (5.22)$$

where the differential-algebraic orders r_{jk} , s_{jk} are given in Tables 5.3 and 5.5.

Table 5.3. orders r_{jk} .

1
1 2
1 2 3
1 2 3 4
1 2 3 4 4
1 2 3 4 4 5
1 2 3 4 4 5 5
1 2 3 4 4 5 6 5
1 2 3 4 4 5 6 6 5
1 2 3 4 4 5 6 7 6 5

Table 5.4. orders s_{jk} .

2
2 2
2 2 3
2 2 3 4
2 2 3 4 4
2 2 3 4 5 4
2 2 3 4 5 5 4
2 2 3 4 5 6 5 4
2 2 3 4 5 6 6 5 4
2 2 3 4 5 6 7 6 5 4

Proof. We use the expansion (5.11). It follows from $\alpha_i^1 = \beta_i^1 = 0$ (for all $i \geq 0$) and from (5.15) that $a_1(x_0) = b_1(x_0) = 0$. Since $a_j(x)$ and $b_j(x)$ are smooth functions we obtain $a_1(x_0 + H) = \mathcal{O}(H)$, $b_1(x_0 + H) = \mathcal{O}(H)$ and the errors of Y_{j1} , Z_{j1} are seen to be of size $\mathcal{O}(H^2)$. This verifies the entries of the first columns of Tables 5.3 and 5.4. In the same way we deduce that $a_2(x_0 + H) = \mathcal{O}(H)$. However, since $\beta_0^2 \neq 0$ in general, we have $b_2(x_0) \neq 0$ by (5.15) and the term $b_2(x_0 + H)$ is only of size $\mathcal{O}(1)$. One extrapolation of the numerical solution eliminates the terms with $j = 1$ in (5.11). The error is thus of size $\mathcal{O}(H^3)$ for Y_{j2} but only $\mathcal{O}(H^2)$ for Z_{j2} , verifying the second columns of Tables 5.3 and 5.4. If we continue the extrapolation process, the smooth parts of the error expansion (5.11) are eliminated one after the other. The perturbation terms, however, are *not* eliminated.

For the y -component the first non-vanishing perturbation for $i \geq n_1 = 1$ is α_1^6 . Therefore, the diagonal elements of the extrapolation tableau for the y -component (Table 5.3) contain an error term of size $\mathcal{O}(H^6)$ (observe that α_1^6 is multiplied by h^6 in (5.11)). The elements $Y_{j,j-1}$ of the first subdiagonal depend only on $Y_{\ell 1} = y_{n_\ell}$ for $\ell \geq 2$. Since $n_2 \geq 2$, only the perturbations α_i^j with $i \geq 2$ can have an influence. We see from (5.12) that the first non-vanishing perturbation for $i \geq 2$ is α_2^7 . This explains the $\mathcal{O}(H^7)$ error term in the first subdiagonal of Table 5.3.

For the z -component, β_1^4 is the first perturbation term for $i \geq 1$. Hence the diagonal entries of the extrapolation tableau for the z -component contain an error of size $\mathcal{O}(H^4)$. All other entries of Tables 5.3 and 5.4 can be verified analogously. \square

If we consider a step number sequence $\{n_j\}$ which is different from the harmonic sequence, we obtain the corresponding order tableaux as follows: the j th diagonal of the new tableau is the n_j th diagonal of Table 5.3 and 5.4, respectively.

Theorem 5.4 then remains valid with r_{jk} , s_{jk} given by these new tableaux. This implies that a larger n_1 , say $n_1 = 2$ increases, the order of the extrapolated values. Numerical computations have shown that the sequence

$$\{2, 3, 4, 5, 6, \dots\} \quad (5.23)$$

is superior to the harmonic sequence. It is therefore recommended for SEULEX.

It is interesting to study the influence of the perturbation terms on the extrapolated values. Suppose that $\alpha_{n_1}^j$ (or $\beta_{n_1}^j$) is the leading perturbation term in Y_{11} (or Z_{11}). Because of the recursion (IV.9.26) all Y_{kk} then contain an error term of the form $C_k H^j \alpha_{n_1}^j$, whereas the Y_{jk} (for $j > k$) do not depend on $\alpha_{n_1}^j$. The error constants C_k are given recursively by

$$C_1 = \frac{1}{n_1^j}, \quad C_k = -\frac{n_1}{n_k - n_1} C_{k-1} \quad (5.24)$$

and tend to zero exponentially, if k increases.

Error Expansion for Singular Perturbation Problems

Our aim is to extend the analysis of Example 5.1 to general singular perturbation problems

$$\begin{aligned} y' &= f(y, z), & y(0) &= y_0 \\ \varepsilon z' &= g(y, z), & z(0) &= z_0, \quad 0 < \varepsilon \ll 1, \end{aligned} \quad (5.25)$$

where the solution $y(x)$, $z(x)$ is assumed to be sufficiently smooth (i.e., its derivatives up to a certain order are bounded independently of ε). An important observation in Example 5.1 was the existence of smooth solutions of the (linear) differential equations for the coefficients $b_i(x)$. In the general situation we shall be concerned with equations of the form

$$\begin{aligned} a' &= f_y(x)a + f_z(x)b + c(x, \varepsilon) \\ \varepsilon b' &= g_y(x)a + g_z(x)b + d(x, \varepsilon) \end{aligned} \quad (5.26)$$

(the coefficients $f_y(x) = f_y(y(x), z(x))$, etc. depend smoothly on ε because the solution of (5.25) itself depends on ε , even if f and g are ε -independent).

Lemma 5.5. *Suppose that the logarithmic norm of $g_z(x)$ satisfies*

$$\mu(g_z(x)) \leq -1 \quad \text{for } 0 \leq x \leq \bar{x}. \quad (5.27)$$

For a given value

$$a(0) = a_0^0 + \varepsilon a_0^1 + \dots + \varepsilon^N a_0^N + \mathcal{O}(\varepsilon^{N+1})$$

there exists a unique (up to $\mathcal{O}(\varepsilon^{N+1})$)

$$b(0) = b_0^0 + \varepsilon b_0^1 + \dots + \varepsilon^N b_0^N + \mathcal{O}(\varepsilon^{N+1})$$

such that the solutions $a(x)$, $b(x)$ of (5.26) and their first N derivatives are bounded independently of ε .

Proof. We insert the finite expansions

$$\widehat{a}(x) = \sum_{i=0}^N \varepsilon^i a_i(x), \quad \widehat{b}(x) = \sum_{i=0}^N \varepsilon^i b_i(x)$$

with ε -independent coefficients $a_i(x)$, $b_i(x)$ into (5.26) and compare powers of ε (see Section VI.2). This leads to the differential-algebraic system (2.4). Consequently, a_0^0 determines b_0^0 ; these two together with a_0^1 determine b_0^1 , etc. The remainders $a(x) - \widehat{a}(x)$, $b(x) - \widehat{b}(x)$ are then estimated as in the proof of Theorem 2.1. \square

The next result exhibits the dominant perturbation terms in an asymptotic expansion of the error of the linearly implicit Euler method, when it is applied to a singular perturbation problem.

Theorem 5.6 (Hairer & Lubich 1988). *Assume that the solution of (5.25) is smooth. Under the condition*

$$\|(I - \gamma g_z(0))^{-1}\| \leq \frac{1}{1 + \gamma} \quad \text{for all } \gamma \geq 1 \quad (5.28)$$

(which is a consequence of (5.27) and Theorem IV.11.2), the numerical solution of (5.1) possesses for $\varepsilon \leq h$ a perturbed asymptotic expansion of the form

$$y_i = y(x_i) + h a_1(x_i) + h^2 a_2(x_i) + \mathcal{O}(h^3) \quad (5.29)$$

$$- \varepsilon f_z(0) g_z^{-1}(0) \left(I - \frac{h}{\varepsilon} g_z(0) \right)^{-i} (h b_1(0) + h^2 b_2(0))$$

$$z_i = z(x_i) + h b_1(x_i) + h^2 b_2(x_i) + \mathcal{O}(h^3) \quad (5.30)$$

$$- \left(I - \frac{h}{\varepsilon} g_z(0) \right)^{-i} (h b_1(0) + h^2 b_2(0))$$

where $x_i = ih \leq H$ with H sufficiently small (but independent of ε). The smooth functions $a_j(x)$, $b_j(x)$ satisfy

$$a_1(0) = \mathcal{O}(\varepsilon^2), \quad a_2(0) = \mathcal{O}(\varepsilon), \quad b_1(0) = \mathcal{O}(\varepsilon), \quad b_2(0) = \mathcal{O}(1).$$

Proof. This proof is organized like that of Theorem 5.3. In part (a) we recursively construct truncated expansions (for $M \leq 2$)

$$\begin{aligned} \widehat{y}_i &= y(x_i) + \sum_{j=1}^M h^j (a_j(x_i) + \alpha_i^j) \\ \widehat{z}_i &= z(x_i) + \sum_{j=1}^M h^j (b_j(x_i) + \beta_i^j) \end{aligned} \quad (5.31)$$

such that

$$\begin{pmatrix} I - hf_y(0) & -hf_z(0) \\ -hg_y(0) & \varepsilon I - hg_z(0) \end{pmatrix} \begin{pmatrix} \widehat{y}_{i+1} - \widehat{y}_i \\ \widehat{z}_{i+1} - \widehat{z}_i \end{pmatrix} = h \begin{pmatrix} f(\widehat{y}_i, \widehat{z}_i) \\ g(\widehat{y}_i, \widehat{z}_i) \end{pmatrix} + \mathcal{O}(h^{M+2}). \quad (5.32)$$

The smooth functions $a_j(x)$, $b_j(x)$ clearly depend on ε , but are independent of h . The perturbation terms α_i^j , β_i^j (for $i \geq 1$), however, will depend smoothly on ε and on ε/h . As in the case $\varepsilon = 0$, we shall require that (5.15) and (5.16) hold. The differences $y_i - \widehat{y}_i$ and $z_i - \widehat{z}_i$ will then be estimated in part b).

a) The case $M = 0$ is obvious. Indeed, the values $\widehat{y}_i = y(x_i)$, $\widehat{z}_i = z(x_i)$ satisfy (5.32) with $M = 0$. The construction of the coefficients in (5.31) is done in two steps.

First step ($M = 1$). We insert (5.31) into (5.32) and compare the smooth coefficients of h^2 . This gives

$$a_1'(x) + \frac{1}{2}y''(x) - f_y(0)y'(x) - f_z(0)z'(x) = f_y(x)a_1(x) + f_z(x)b_1(x) \quad (5.33a)$$

$$\varepsilon b_1'(x) + \frac{\varepsilon}{2}z''(x) - g_y(0)y'(x) - g_z(0)z'(x) = g_y(x)a_1(x) + g_z(x)b_1(x) \quad (5.33b)$$

By Lemma 5.5 the initial value $b_1(0)$ is uniquely determined by $a_1(0)$. Differentiation of $\varepsilon z' = g(y, z)$ with respect to x gives $\varepsilon z''(x) = g_y(x)y'(x) + g_z(x)z'(x)$. Inserted into (5.33b) this yields the relation

$$g_y(0)a_1(0) + g_z(0)b_1(0) = \mathcal{O}(\varepsilon) \quad (5.34)$$

with known right-hand side.

As to the perturbation terms, we obtain by collecting everything up to $\mathcal{O}(h^2)$

$$\begin{aligned} \alpha_{i+1}^1 - \alpha_i^1 - hf_y(0)(\alpha_{i+1}^1 - \alpha_i^1) - hf_z(0)(\beta_{i+1}^1 - \beta_i^1) \\ = hf_y(x_i)\alpha_i^1 + hf_z(x_i)\beta_i^1 \\ \varepsilon(\beta_{i+1}^1 - \beta_i^1) - hg_y(0)(\alpha_{i+1}^1 - \alpha_i^1) - hg_z(0)(\beta_{i+1}^1 - \beta_i^1) \\ = hg_y(x_i)\alpha_i^1 + hg_z(x_i)\beta_i^1 \end{aligned} \quad (5.35)$$

and try to determine the most important parts of this. We firstly replace $hf_y(x_i)\alpha_i^1$ by $hf_y(0)\alpha_i^1$ and similarly for three other terms. This is motivated by the fact that we search for exponentially decaying α_i . Therefore with $x_i = ih$,

$$(f_y(x_i) - f_y(0))\alpha_i^1 = \mathcal{O}(h).$$

Then many terms cancel in (5.35). We next observe that $\beta_{i+1}^1 - \beta_i^1$ is multiplied by ε , but not $\alpha_{i+1}^1 - \alpha_i^1$. This suggests that the β_{i+1}^1 are an order of magnitude larger than α_{i+1}^1 . Neglecting therefore α_{i+1}^1 where it competes with β_{i+1}^1 , we are led to define

$$\alpha_{i+1}^1 - \alpha_i^1 = hf_z(0)\beta_{i+1}^1 \quad (5.33c)$$

$$\varepsilon(\beta_{i+1}^1 - \beta_i^1) = hg_z(0)\beta_{i+1}^1. \quad (5.33d)$$

It remains to verify a posteriori, that there exist solutions of (5.33a,b,c,d) which produce an error term $\mathcal{O}(h^3)$ in (5.32): from (5.33d) we obtain

$$\beta_i^1 = \left(I - \frac{h}{\varepsilon} g_z(0) \right)^{-i} \beta_0^1. \quad (5.36a)$$

Since we require $\alpha_i^1 \rightarrow 0$ for $i \rightarrow \infty$, the solution of (5.33c) is given by

$$\alpha_i^1 = \varepsilon f_z(0) g_z^{-1}(0) \left(I - \frac{h}{\varepsilon} g_z(0) \right)^{-i} \beta_0^1. \quad (5.36b)$$

For $i = 0$ this implies the relation

$$\alpha_0^1 = \varepsilon f_z(0) g_z^{-1}(0) \beta_0^1. \quad (5.37)$$

The assumption (5.15) together with (5.34) and (5.37) uniquely determine the coefficients $a_1(0)$, $b_1(0)$, α_0^1 , β_0^1 . We remark that $b_1(0) = \mathcal{O}(\varepsilon)$ and $a_1(0) = \mathcal{O}(\varepsilon^2)$. Using the fact that $\alpha_i^1 = \mathcal{O}(\varepsilon^2)$ and $\varepsilon \leq h$, one easily verifies that the quantities (5.31) with $M = 1$ satisfy (5.32).

Second step ($M = 2$). Comparing the smooth coefficients of h^3 in (5.32) gives two differential equations for $a_2(x)$, $b_2(x)$ which are of the form (5.26). It follows from Lemma 5.5 that the initial values have to satisfy a relation

$$g_y(0)a_2(0) + g_z(0)b_2(0) = \mathcal{O}(1) \quad (5.38)$$

with known right-hand side. As in the first step we require for the perturbations

$$\begin{aligned} \alpha_{i+1}^2 - \alpha_i^2 &= h f_z(0) \beta_{i+1}^2 \\ \varepsilon(\beta_{i+1}^2 - \beta_i^2) &= h g_z(0) \beta_{i+1}^2. \end{aligned} \quad (5.39)$$

and obtain the formulas (5.36) and (5.37) with α_i^1 , β_i^1 replaced by α_i^2 , β_i^2 . Again the values $a_2(0)$, $b_2(0)$, α_0^2 , β_0^2 are uniquely determined by (5.15), (5.38), and (5.37). Due to the $\mathcal{O}(1)$ term in (5.38) we only have $b_2(0) = \mathcal{O}(1)$ and $a_2(0) = \mathcal{O}(\varepsilon)$.

We still have to verify (5.32) with $M = 2$. In the left-hand side we have neglected terms of the form $h f_y(0)(h \alpha_i^1 + h^2 \alpha_i^2)$. This is justified, because $\alpha_i^1 = \mathcal{O}(\varepsilon^2)$, $\alpha_i^2 = \mathcal{O}(\varepsilon)$ and $\varepsilon \leq h$. The most dangerous term, neglected in the right-hand side of (5.32) is

$$h(f_z(x_i) - f_z(0))(h \beta_i^1 + h^2 \beta_i^2). \quad (5.40)$$

However, $f_z(x_i) - f_z(0) = \mathcal{O}(ih)$, and $\beta_i^1 = \mathcal{O}(\varepsilon 2^{-i})$, $\beta_i^2 = \mathcal{O}(2^{-i})$ by (5.28) and $\varepsilon \leq h$. This shows that the term (5.40) is also of size $\mathcal{O}(h^4)$, so that (5.32) holds with $M = 2$.

b) In order to estimate the remainder term, i.e., the differences $\Delta y_i = y_i - \widehat{y}_i$, $\Delta z_i = z_i - \widehat{z}_i$ we subtract (5.32) from (5.1) and eliminate Δy_{i+1} and Δz_{i+1} . This

gives

$$\begin{pmatrix} \Delta y_{i+1} \\ \Delta z_{i+1} \end{pmatrix} = \begin{pmatrix} \Delta y_i \\ \Delta z_i \end{pmatrix} + \begin{pmatrix} I + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & (\frac{\varepsilon}{h}I - g_z(0))^{-1} \end{pmatrix} \begin{pmatrix} h(f(y_i, z_i) - f(\hat{y}_i, \hat{z}_i)) \\ g(y_i, z_i) - g(\hat{y}_i, \hat{z}_i) \end{pmatrix} + \begin{pmatrix} \mathcal{O}(h^{M+2}) \\ \mathcal{O}(h^{M+1}) \end{pmatrix}.$$

Due to (5.28) and $\varepsilon \leq h$ we have

$$\left\| I + \left(\frac{\varepsilon}{h}I - g_z(0) \right)^{-1} g_z(0) \right\| = \left\| \left(I - \frac{h}{\varepsilon} g_z(0) \right)^{-1} \right\| \leq \frac{\varepsilon}{\varepsilon + h} \leq \frac{1}{2}. \quad (5.41)$$

We therefore again obtain (5.20) with some $|\varrho| < 1$, if H is sufficiently small. We then deduce the result as in the proof of Theorem 5.3. \square

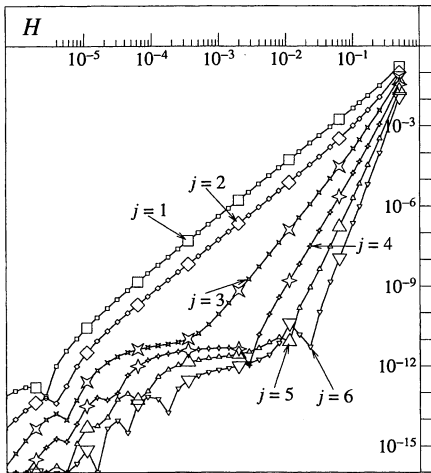


Fig. 5.1. Step size/error diagram

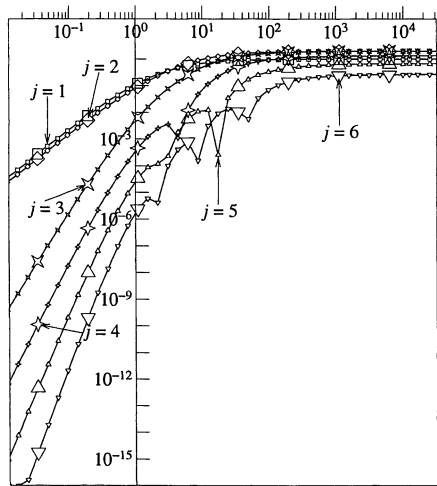


Fig. 5.2. T_{jj} as functions of H/ε

Of course, it is possible to add a third step to the above proof. However, the recursions for α_i^3 , β_i^3 are no longer as simple as in (5.33) or (5.39). Moreover, the perturbations of (5.29) and (5.30) already describe very well the situation encountered in practice. We shall illustrate this with the following numerical example (see also Hairer & Lubich 1988).

Consider van der Pol's equation (2.73) with $\varepsilon = 10^{-5}$ and with initial values (2.74) on the smooth solution. We take the step number sequence (5.23) and apply Method (5.1) n_j times with step size $h = H/n_j$. The numerical result Y_{j1} , Z_{j1} is then extrapolated according to (IV.9.26). In Fig. 5.1 we show in logarithmic scale the errors $|Z_{jj} - z(H)|$ for $j = 1, 2, \dots, 6$ as functions of H . We observe that whenever the error is larger than $\varepsilon^2 = 10^{-10}$, the curves appear as straight lines with slopes 2, 2, 3, 4, 5, and 6, respectively. If its slope is q , we have $\log(\text{error}) \approx$

$q \log H + \text{Const}$, or equivalently $\text{error} \approx CH^q$. This corresponds (with exception of the last one) to the orders predicted by the subdiagonal entries of Table 5.4 for the case $\varepsilon = 0$.

In order to understand the irregular behaviour of the curves when the error becomes smaller than $\varepsilon^2 = 10^{-10}$, we study the influence of extrapolation on the perturbation terms in (5.30). Since $b_1(0)$ contains a factor ε , the dominant part of the perturbation in Z_{j1} is $(I - (h/\varepsilon)g_z(0))^{-n_j} h^2 b_2(0)$, where $b_2(0)$ is some constant and $h = H/n_j$. We assume the matrix $g_z(0)$ to be diagonalized and put $g_z(0) = -1$. The dominant perturbation in Z_{j1} is therefore $\varepsilon^2 T_{j1} b_2(0)$, where

$$T_{j1} = \left(\frac{H}{\varepsilon n_j} \right)^2 \left(1 + \frac{H}{\varepsilon n_j} \right)^{-n_j}. \quad (5.42)$$

Due to the linearity of extrapolation, the dominant perturbation in Z_{jj} will be $\varepsilon^2 T_{jj} b_2(0)$, where T_{jj} is obtained from (5.42) and (IV.9.26). For the step number sequence (5.23) the values of T_{jj} are plotted as functions of H/ε in Fig. 5.2. For large values of H/ε the curves appear as horizontal lines. This is a consequence of our choice $n_1 = 2$ and of the fact that

$$T_{jj} = C_j \cdot \left(\frac{H}{\varepsilon} \right)^{2-n_1} + \mathcal{O}\left(\left(\frac{H}{\varepsilon} \right) \right)^{1-n_1} \quad \text{for } \frac{H}{\varepsilon} \rightarrow \infty,$$

where $C_1 = 1$ and the other C_j are given by the recursion (5.24).

The errors of Fig. 5.1 are now seen to be a superposition of the errors, predicted from the case $\varepsilon = 0$ (Theorem 5.4), and of the perturbations of Fig. 5.2 scaled by a factor $\mathcal{O}(\varepsilon^2)$.

Remark. As mentioned in Sect. VI.1, the *implicit Euler* discretization possesses a classical asymptotic expansion for differential-algebraic problems (1.6) of index 1 (case $\varepsilon = 0$). However, for singular perturbation problems, perturbations of the same type as in (5.29) and (5.30) are present. The only difference is that all $b_i(0)$ contain a factor ε for the implicit Euler method. For details and numerical experiments we refer to Hairer & Lubich (1988). A related analysis for a slightly different class of singular perturbation problems is presented in Auzinger, Frank & Macsek (1990).

Dense Output

Extrapolation methods typically take very large (basic) step sizes during integration. This makes it important that the method possess a continuous numerical solution. The first attempt to get a dense output for extrapolation methods is due to Lindberg (1972). His approach, however, imposes severe restrictions on the step number sequence. We present here the dense output of Hairer & Ostermann (1990), which exists for any step number sequence.

The main idea (due to Ch. Lubich) is the following: when computing the j -th entry of the extrapolation tableau, we consider not only $Y_{j1} = y_{n_j}$, but also

compute the difference $(y_{n_j} - y_{n_j-1})/h_j$. Since these expressions possess an h -expansion, their extrapolation gives an accurate approximation to $y'(x_0 + H)$. By considering higher differences, we get also approximations to higher derivatives of $y(x)$ at $x_0 + H$. They are then used for Hermite interpolation. The reason for computing the derivatives only at the right end of the basic interval, is the presence of perturbation terms as described in Theorems 5.3 and 5.6. These perturbations are large at the beginning (near the initial value), but decrease exponentially for increasing i . For the same reason, one must not use differences of a too high an order. We thus choose an integer λ (usually 0 or 1) and avoid the values $y_0, \dots, y_{n_1+\lambda-2}$ for the computation of the finite differences. We remark that a similar idea was used by Deufilhard & Nowak (1987) to construct consistent initial values for differential-algebraic problems.

An algorithmic description of the dense output for the linearly implicit Euler method is as follows (we suppose that the value $Y_{\kappa\kappa}$ has been accepted as a numerical approximation to $y(x_0 + H)$).

Step 1. For each $j \in \{1, \dots, \kappa\}$ we compute

$$r_j^{(k)} = \frac{\nabla^k y_{n_j}^{(j)}}{h_j^k} \quad \text{for } k = 1, \dots, j - \lambda. \quad (5.43)$$

Here $y_i^{(j)}$ is the approximation of $y(x_i)$, obtained during the computation of Y_{j1} , and $\nabla y_i = y_i - y_{i-1}$ is the backward difference operator.

Step 2. We extrapolate $r_j^{(k)}$, $(\kappa - k - \lambda)$ times. This yields the improved approximation $r^{(k)}$ to $y^{(k)}(x_0 + H)$.

Step 3. We define the polynomial $P(\theta)$ of degree κ by

$$\begin{aligned} P(0) &= y_0, & P(1) &= Y_{\kappa\kappa} \\ P^{(k)}(1) &= H^k r^{(k)} & \text{for } k &= 1, \dots, \kappa - 1. \end{aligned} \quad (5.44)$$

The following theorem shows to which order these polynomials approximate the exact solution.

Theorem 5.7 (Hairer & Ostermann 1990). *Consider a nonstiff differential equation and let $\lambda \in \{0, 1\}$. Then, the error of the interpolation polynomial $P(\theta)$ satisfies*

$$P(\theta) - y(x_0 + \theta H) = \mathcal{O}(H^{\kappa+1-\lambda}) \quad \text{for } H \rightarrow 0.$$

Proof. Since $P(\theta)$ is a polynomial of degree κ , the error due to interpolation is of size $\mathcal{O}(H^{\kappa+1})$. We know that $Y_{\kappa\kappa} - y(x_0 + H) = \mathcal{O}(H^{\kappa+1})$. Therefore it suffices to prove that

$$r^{(k)} = y^{(k)}(x_0 + H) + \mathcal{O}(H^{\kappa-k-\lambda+1}) \quad \text{for } k = 1, \dots, \kappa - 1. \quad (5.45)$$

Due to the asymptotic expansion of the global error $y_i - y(x_i)$, the approximations $r_j^{(k)}$ also have an expansion of the form

$$r_j^{(k)} = y^{(k)}(x_0 + H) + h_j a_1^{(k)} + h_j^2 a_2^{(k)} + \dots \quad (5.46)$$

The statement (5.45) now follows from the fact that each extrapolation eliminates one power of h in (5.46). \square

It is now natural to investigate the error of the dense output $P(\theta)$ also for stiff differential equations, such as singular perturbation problems. We shall treat here the limit case $\varepsilon = 0$ which is easier to analyse and, nevertheless, gives much insight into the structure of the error for very stiff problems.

For the differential-algebraic system (1.6) one defines the dense output in exactly the same way as for ordinary differential equations. As the system (1.6) is partitioned into y - and z -components, it is convenient to denote the corresponding interpolation polynomials by $P(\theta)$ and $Q(\theta)$, respectively.

Theorem 5.8 (Hairer & Ostermann 1990). *Let $y(x)$, $z(x)$ be the solution of (1.6). Suppose that the step number sequence satisfies $n_1 + \lambda \geq 2$ with $\lambda \in \{0, 1\}$. We then have*

$$\begin{aligned} P(\theta) - y(x_0 + \theta H) &= \mathcal{O}(H^{\kappa+1-\lambda}) + \mathcal{O}(H^{r+1}), \\ Q(\theta) - z(x_0 + \theta H) &= \mathcal{O}(H^{\kappa+1-\lambda}) + \mathcal{O}(H^s), \end{aligned} \quad (5.47)$$

where r and s are the $(\kappa + n_1 + \lambda - 2, \kappa)$ -entries of Table 5.3 and Table 5.4, respectively.

Proof. We use the perturbed asymptotic error expansions of Theorem 5.3. Their smooth terms are treated exactly as in the proof of Theorem 5.7 and yield the $\mathcal{O}(H^{\kappa+1-\lambda})$ error term in (5.47). The second error terms in (5.47) are due to the perturbations in (5.11). We observe that the computation of $r_j^{(k)}$ involves only y_i (or z_i) with $i \geq n_j - j + \lambda$. Since $n_j - j \geq n_1 - 1$, the values $y_0, \dots, y_{n_1+\lambda-2}$ do not enter into the formulas for $r_j^{(k)}$, so that the dominant perturbation comes from $y_{n_1+\lambda-1}$ (or $z_{n_1+\lambda-1}$). \square

It is interesting to note that for $\lambda = 1$, the second error term in (5.47) is of the same size as that in the numerical solution $Y_{\kappa\kappa}$, $Z_{\kappa\kappa}$ (see Theorem 5.4). However, one power of H is lost in the first term of (5.47). On the other hand, one H may be lost in the second error term, if $\lambda = 0$. Both choices lead to a cheap (no additional function evaluations) and accurate dense output. Its order for $\theta \in (0, 1)$ is at most one lower than the order obtained for $\theta = 1$.

Exercises

1. The linearly implicit mid-point rule, applied to the differential-algebraic system (1.6), reads

$$\begin{pmatrix} I - hf_y(0) & -hf_z(0) \\ -hg_y(0) & -hg_z(0) \end{pmatrix} \begin{pmatrix} y_{i+1} - y_i \\ z_{i+1} - z_i \end{pmatrix} \quad (5.48) \\ = - \begin{pmatrix} I + hf_y(0) & hf_z(0) \\ hg_z(0) & hg_z(0) \end{pmatrix} \begin{pmatrix} y_i - y_{i-1} \\ z_i - z_{i-1} \end{pmatrix} + 2h \begin{pmatrix} f(y_i, z_i) \\ g(y_i, z_i) \end{pmatrix}.$$

If we compute y_1, z_1 from (5.2), and if we define the numerical solution at $x_0 + H$ ($H = 2mh$) by

$$y_h(x_0 + H) = \frac{1}{2}(y_{2m+1} + y_{2m-1}), \quad z_h(x_0 + H) = \frac{1}{2}(z_{2m+1} + z_{2m-1}),$$

this algorithm constitutes an extension of (IV.9.16) to differential-algebraic problems.

a) Show that this method integrates the problem (5.8) exactly.

b) Apply the algorithm to

$$y' = 1, \quad 0 = u - y^2, \quad 0 = v - yu, \quad 0 = w - yv, \quad 0 = z - yw$$

with zero initial values and verify the formula

$$\begin{aligned} \frac{1}{2}(z_{2m+1} + z_{2m-1}) - z(x_{2m}) &= -10x_{2m}^3 h^2 + 9x_{2m} h^4 \\ &\quad - (-1)^m \left(\frac{1}{8} x_{2m}^5 + x_{2m}^3 h^2 + 9x_{2m} h^4 \right). \end{aligned}$$

Remark. The error of the z -component thus contains an h -independent term of size $\mathcal{O}(H^5)$, which is not affected by extrapolation.

2. Consider the method of Exercise 1 as the basis of an h^2 -extrapolation method. Prove that for the step number sequence (IV.9.22) the extrapolated values satisfy

$$Y_{jk} - y(x_0 + H) = \mathcal{O}(H^{r_{jk}+1}), \quad Z_{jk} - z(x_0 + H) = \mathcal{O}(H^{s_{jk}})$$

with r_{jk}, s_{jk} given in Tables 5.5 and 5.6.

Hint. Interpret Y_{j1}, Z_{j1} as numerical solution of a Rosenbrock method (Exercise 3 of Sect. IV.9) and verify the order condition derived in Sect. VI.3 (see also Hairer & Lubich (1988b) and C. Schneider (1993)).

Table 5.5. orders r_{jk} .

1
1 3
1 3 5
1 3 5 7
1 3 5 7 7
1 3 5 7 7 7
1 3 5 7 7 7 7

Table 5.6. orders s_{jk} .

2
2 4
2 4 5
2 4 5 5
2 4 5 5 5
2 4 5 5 5 5
2 4 5 5 5 5 5

VI.6 Quasilinear Problems

Quasilinear differential equations are usually understood to be equations in which the highest derivative appears linearly. In the case of first order ODE systems, they are of the form

$$C(y) \cdot y' = f(y), \quad (6.1)$$

where $C(y)$ is a $n \times n$ -matrix. In the regions where $C(y)$ is invertible, Eq. (6.1) can be written as

$$y' = C(y)^{-1} \cdot f(y) \quad (6.1')$$

and every ODE-code can be applied by solving at every function call a linear system. But this would destroy, for example, a banded structure of the Jacobian and it is therefore often preferable to treat Eq. (6.1) directly. If the matrix C is everywhere of rank m ($m < n$), Eq. (6.1) represents a quasilinear differential-algebraic system.

Example: Moving Finite Elements

As an example, we present the classical idea of “moving finite elements”, described in K. Miller & R.N. Miller (1981): the solution $u(x, t)$ of a nonlinear partial differential equation

$$\frac{\partial u}{\partial t} = L(u(x, t)), \quad u(0, t) = u(1, t) = 0, \quad (6.2)$$

where $L(u)$ is an unbounded nonlinear differential operator, is approximated by finite element polygons $v(x, a_1, s_1, \dots, a_n, s_n)$ which satisfy $v(s_j, \dots) = a_j$ (see Fig. 6.1). These polygons form a $2n$ -dimensional manifold in the Hilbert space $L^2(0, 1)$ parametrized by $a_1, s_1, \dots, a_n, s_n$. The idea is now to *move simultaneously* $a(t)$ and $s(t)$ in order to adapt at any time the finite element solution as best as possible to Eq. (6.2).

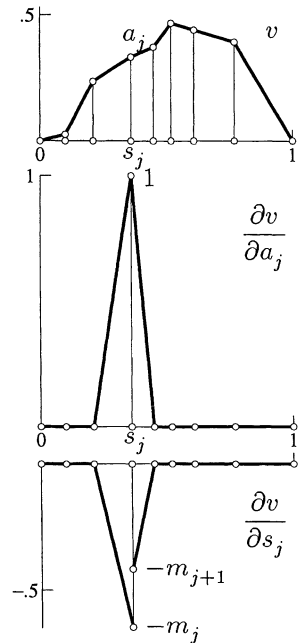


Fig. 6.1. Moving finite elements

We thus require that the defect $\dot{v} - L(v)$ remains always orthogonal to the tangent space. The conditions

$$\int_0^1 (\dot{v} - L(v)) \cdot \frac{\partial v}{\partial a_j} dx = 0 \quad \int_0^1 (\dot{v} - L(v)) \cdot \frac{\partial v}{\partial s_j} dx = 0 \quad (6.3)$$

lead to a system of type (6.1) with

$$\begin{aligned} c_{2j-1,2k-1} &= \int_0^1 \frac{\partial v}{\partial a_j} \cdot \frac{\partial v}{\partial a_k} dx, & c_{2j-1,2k} &= \int_0^1 \frac{\partial v}{\partial a_j} \cdot \frac{\partial v}{\partial s_k} dx, \\ c_{2j,2k-1} &= \int_0^1 \frac{\partial v}{\partial s_j} \cdot \frac{\partial v}{\partial a_k} dx, & c_{2j,2k} &= \int_0^1 \frac{\partial v}{\partial s_j} \cdot \frac{\partial v}{\partial s_k} dx, \\ f_{2j-1} &= \int_0^1 L(v) \cdot \frac{\partial v}{\partial a_j} dx, & f_{2j} &= \int_0^1 L(v) \cdot \frac{\partial v}{\partial s_j} dx. \end{aligned} \quad (6.4)$$

For the partial derivatives of v , sketched in Fig. 6.1, the non-zero of these scalar products become

$$\begin{aligned} c_{2j-1,2j-1} &= \frac{1}{3}(\Delta_j + \Delta_{j+1}) & c_{2j-1,2j} &= -\frac{1}{3}(m_j \Delta_j + m_{j+1} \Delta_{j+1}) \\ c_{2j,2j-1} &= -\frac{1}{3}(m_j \Delta_j + m_{j+1} \Delta_{j+1}) & c_{2j,2j} &= \frac{1}{3}(m_j^2 \Delta_j + m_{j+1}^2 \Delta_{j+1}) + 2\varepsilon^2 \end{aligned} \quad (6.5a)$$

$$\begin{aligned} c_{2j-1,2j+1} &= c_{2j+1,2j-1} = \frac{1}{6}\Delta_{j+1} & c_{2j-1,2j+2} &= c_{2j+2,2j-1} = -\frac{1}{6}m_{j+1}\Delta_{j+1} \\ c_{2j,2j+1} &= c_{2j+1,2j} = -\frac{1}{6}m_{j+1}\Delta_{j+1} & c_{2j,2j+2} &= c_{2j+2,2j} = \frac{1}{6}m_{j+1}^2\Delta_{j+1} - \varepsilon^2 \end{aligned} \quad (6.5b)$$

where

$$\Delta_j = s_j - s_{j-1}, \quad m_j = (a_j - a_{j-1})/\Delta_j, \quad j = 1, \dots, n+1.$$

The matrix $C(y)$ is banded with bandwidth $3 + 1 + 3$. The ε^2 -terms in (6.5) come from an “internodal viscosity” penalty term, explained in Miller & Miller (1981), which aims to regularize the relative movement of the nodes s_j whenever their position is ill-conditioned, which happens to appear in the vicinity of inflection points (see Fig. 6.2).

It is then hoped that the nodes move automatically into the critical regions of the solutions, move with shocks which may appear, and that $a(t)$ and $s(t)$ become smooth functions.

Application to Burgers’ Equation. Burgers’ Equation is given by

$$u_t = -uu_x + \mu u_{xx} \quad \text{or} \quad u_t = -\left(\frac{u^2}{2}\right)_x + \mu u_{xx} \quad (6.6)$$

where $\mu = 1/R$ and R is called the Reynolds number. This is one of the equations originally designed by Burgers (1948) as “a mathematical model illustrating the theory of turbulence”. However, soon afterwards, E. Hopf (1950) presented an analytical solution (see Exercise 1 below) and concluded that “we doubt that Burgers’ equation fully illustrates the statistics of free turbulence. (...) Equation (1) is too simple a model to display chance fluctuations ...”. Nowadays it remains interesting as a nonlinear equation resembling the Navier-Stokes’ equations in fluid dynamics which possesses, for R large, shock waves and, for $R \rightarrow \infty$,

discontinuous solutions. Here, the integrals in (6.4) become

$$f_{2j-1} = A_j + \mu B_j, \quad f_{2j} = C_j + \mu D_j, \quad j = 1, \dots, n. \quad (6.5c)$$

where

$$\begin{aligned} A_j &= (a_{j-1} - a_j) \left(\frac{1}{3} a_j + \frac{1}{6} a_{j-1} \right) + (a_j - a_{j+1}) \left(\frac{1}{3} a_j + \frac{1}{6} a_{j+1} \right), \\ B_j &= (m_{j+1} - m_j), \\ C_j &= -m_j (a_{j-1} - a_j) \left(\frac{1}{3} a_j + \frac{1}{6} a_{j-1} \right) - m_{j+1} (a_j - a_{j+1}) \left(\frac{1}{3} a_j + \frac{1}{6} a_{j+1} \right), \\ D_j &= (m_{j+1} - m_j) \left(\frac{1}{2} m_{j+1} + \frac{1}{2} m_j \right), \end{aligned} \quad (6.5d)$$

(in the case of D_j appears the product of a Dirac δ function with a discontinuous function; these must be suitably "mollified"). We choose as initial function

$$u(x, 0) = (\sin(3\pi x))^2 \cdot (1-x)^{3/2}, \quad \mu = 0.0003 \quad (6.7)$$

and as initial positions

$$s_j = j/(n+1), \quad a_j = u(s_j, 0), \quad j = 1, \dots, n, \quad n = 100,$$

and solve the problem with smoothing parameter $\varepsilon = 10^{-2}$ for $0 \leq t \leq 1.9$. Two shock waves arise which later fuse into one (see Fig. 6.2).

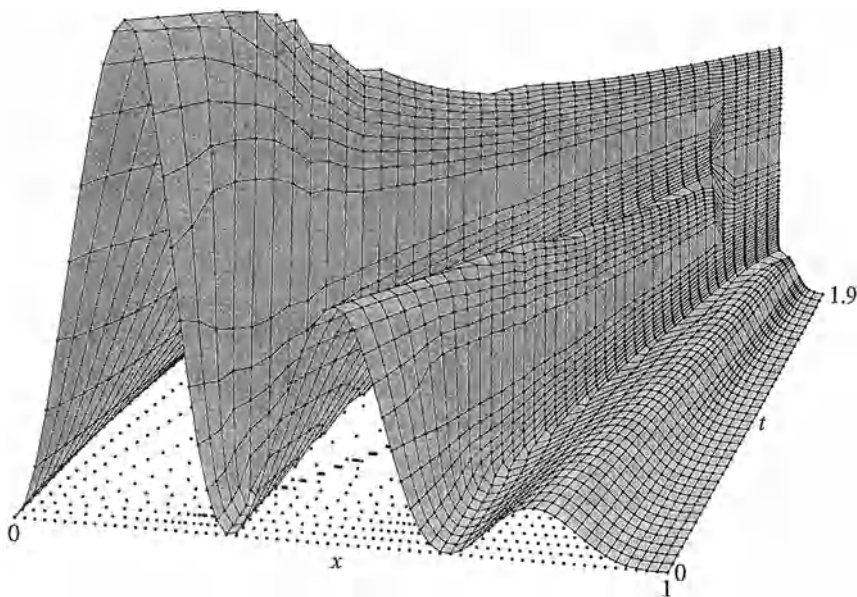


Fig. 6.2. Moving finite element solution of Burgers' equation

Problems of Index One

For invertible $C(y)$, Eq. (6.1) is an ordinary differential equation and standard theory (for existence and uniqueness results) can be applied. If the matrix is ill-conditioned or even singular, new investigations are necessary. In order to exclude equations with singularities, such as $xy' = (q + bx)y$ (see Sect. I.5), we assume that

$$C(y) \text{ has constant rank } m \ (m < n) \quad (6.8)$$

in a neighbourhood of the solution. Then the columns of $C(y)$ span an m -dimensional subspace $\mathcal{I}m C(y)$ which moves with y . Clearly, in order that (6.1) can make sense, we need consistent initial values, i.e., we need

$$f(y_0) \in \mathcal{I}m C(y_0). \quad (6.9)$$

We shall now show, how, under a certain condition, this property can be satisfied for all x and determines uniquely the solution: choose a nonsingular matrix

$$T(y) = \begin{pmatrix} T_1(y) \\ T_2(y) \end{pmatrix} \quad \text{such that} \quad T(y)C(y) = \begin{pmatrix} B_1(y) \\ 0 \end{pmatrix}; \quad (6.10)$$

this means that the rows of $T_2(y)$ must span the $(n - m)$ -dimensional orthogonal complement of $\mathcal{I}m C(y)$. Then we multiply Eq. (6.1) by $T(y)$ and obtain

$$\begin{pmatrix} B_1(y) \\ 0 \end{pmatrix} y' = \begin{pmatrix} T_1(y)f(y) \\ T_2(y)f(y) \end{pmatrix}, \quad (6.11)$$

so that the condition corresponding to (6.9) becomes visible in the form $T_2(y)f(y) = 0$. Differentiating this relation and inserting the derivative into the second part of (6.11), we obtain

$$\begin{pmatrix} B_1(y) \\ (T_2 f)'(y) \end{pmatrix} y' = \begin{pmatrix} (T_1 f)(y) \\ 0 \end{pmatrix}, \quad (6.12)$$

which is a *regular* quasilinear equation if the matrix

$$\begin{pmatrix} B_1(y) \\ (T_2 f)'(y) \end{pmatrix} \quad \text{is invertible.} \quad (6.13)$$

Lemma 6.1. *Let the matrix $C(y)$ satisfy (6.8) and (6.13), and let the initial values y_0 fulfill (6.9). Then, the quasilinear problem $C(y)y' = f(y)$, $y(x_0) = y_0$ possesses a locally unique solution.*

Proof. Condition (6.9) means that $T_2(y_0)f(y_0) = 0$ and the second part of (6.12) assures that $(T_2(y(x))f(y(x)))' = 0$. Therefore we have $(T_2 f)(y(x)) = 0$ for all x , and the solution of (6.12) solves also (6.11) and (6.1). \square

The following result gives a consequence of condition (6.13) which shall be essential in the later discussions of feasibility of numerical procedures.

Lemma 6.2. *Assume that $C(y)$ satisfies (6.8) and (6.13). If $f(y_0) = C(y_0)y'_0$, then the matrix*

$$C(y) + \lambda(f'(y_0) - \Gamma(y_0, y'_0))$$

is invertible for sufficiently small $\lambda \neq 0$ and for y sufficiently close to y_0 . Here,

$$\Gamma(y, y') = \frac{\partial}{\partial y}(C(y)y').$$

Proof. Condition (6.13) implies that

$$T(y)C(y) + \lambda(Tf)'(y_0) \quad \text{is invertible} \quad (6.14)$$

for small $\lambda \neq 0$ and y close to y_0 . Using $T'C + TC' = B'$ we have

$$(T'f)(y_0) = T'(y_0)C(y_0)y'_0 = -T(y_0)\Gamma(y_0, y'_0) + B'(y_0)y'_0. \quad (6.15)$$

Since $B'(y_0)y'_0$ does not contribute to the lower block of the matrix (6.14), it can be neglected after insertion of $(Tf)' = Tf' + T'f$ and (6.15) into (6.14). This implies that

$$T(y)C(y) + \lambda T(y_0)(f'(y_0) - \Gamma(y_0, y'_0)) \quad \text{is invertible.}$$

The statement of the Lemma now follows from a continuity argument. \square

Numerical Treatment of $C(y)y' = f(y)$

As has been said above, in the case of invertible matrices $C(y)$, one can eventually apply an explicit numerical method to (6.1'). However, if Eq. (6.1') is stiff, implicit methods have to be applied. In this case it may be advantageous to have methods that avoid the computation of the Jacobian of $C(y)^{-1}f(y)$.

Transformation to Semi-Explicit Form. In the case where (6.1') is stiff or where $C(y)$ is singular and satisfies (6.8) and (6.13) we introduce $z = y'$ as new variable, such that system (6.1) becomes of the semi-explicit form

$$\begin{aligned} y' &= z \\ 0 &= C(y)z - f(y) \end{aligned} \quad (6.16)$$

Here, all methods of the preceding sections can be applied (at least formally). The study of convergence, however, needs further investigation, because Condition (1.7) is no longer satisfied here.

Implicit Runge-Kutta and Multistep Methods. With the ε -embedding approach (see (1.11) for Runge-Kutta methods and (2.2) for multistep methods) we are led to

nonlinear equations, which, when solved by Newton iterations, require the solution of linear systems of the form

$$\begin{pmatrix} -\alpha I & I \\ \Gamma(y_0, z_0) - f'(y_0) & C(y_0) \end{pmatrix}. \quad (6.17)$$

Here $\alpha = (\gamma h)^{-1}$, and γ is an eigenvalue of the Runge-Kutta matrix. By Lemma 6.2 this matrix is invertible for small enough $h > 0$. Convergence follows from the results of Sections VII.3 and VII.4 (see Exercise 2).

Rosenbrock Methods. Method (4.3) applied to system (6.16) leads to

$$\begin{pmatrix} v_i \\ w_i \end{pmatrix} = \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} + \sum_{j=1}^{i-1} \alpha_{ij} \begin{pmatrix} k_j \\ \ell_j \end{pmatrix}, \quad \begin{pmatrix} y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} + \sum_{i=1}^s b_i \begin{pmatrix} k_i \\ \ell_i \end{pmatrix}. \quad (6.18a)$$

$$\begin{pmatrix} k_i \\ 0 \end{pmatrix} = h \begin{pmatrix} w_i \\ C(v_i)w_i - f(v_i) \end{pmatrix} + h \begin{pmatrix} 0 & I \\ \Gamma_0 - f'_0 & C(y_0) \end{pmatrix} \sum_{j=1}^i \gamma_{ij} \begin{pmatrix} k_j \\ \ell_j \end{pmatrix}. \quad (6.18b)$$

Again, it can be seen that (6.18b) represents a linear system whose regularity is assured by Lemma 6.2. However, since Condition (1.7) is not satisfied, a new theory for the order conditions of the local error as well as for convergence of the global error is necessary. This theory reveals, for example, that new order conditions for the coefficients are necessary and explains why, say, the code RODAS, directly applied to (6.16), does not give precise results. For full details we refer the reader to the original publication Lubich & Roche (1990).

Extrapolation Methods

The first problem is to find suitable linearly implicit Euler discretizations for (6.1), to serve as basic method for the extrapolation algorithm (see Sect. IV.9).

Method of Deuffhard & Nowak. Applying the linearly implicit Euler method (IV.9.15) to the differential equation (6.1') we obtain

$$(I - hA)(y_{i+1} - y_i) = hC(y_i)^{-1}f(y_i) \quad (6.19)$$

where

$$A \approx (C^{-1}f)'(y_0) = C(y_0)^{-1}(f'(y_0) - \Gamma(y_0, y'_0))$$

with $\Gamma(y, y')$ as in Lemma 6.2. Multiplication of (6.19) with $C(y_i)$ yields

$$(C(y_i) - hC(y_i)C(y_0)^{-1}J)(y_{i+1} - y_i) = hf(y_i)$$

with $J = f'(y_0) - \Gamma(y_0, y'_0)$. Deuffhard & Nowak (1987) suggest to replace $C(y_i)C(y_0)^{-1}$ by the identity matrix, which “may be interpreted as just introducing an approximation error into the Jacobian matrix”. This leads to the discretization

$$(C(y_i) - hJ)(y_{i+1} - y_i) = hf(y_i) \quad (6.20)$$

which represents the basic step for the code LIMEX described in Deuffhard & Nowak (1987). The regularity of the matrix of this linear system is again assured by Lemma 6.2.

The computation of J requires an approximation to $z_0 = y'_0$. Such consistent initial values must be computed explicitly for the first basic steps, and are obtained by extrapolation of

$$z_n = (y_n - y_{n-1})/h \quad (6.21)$$

in the subsequent steps.

Linearly-Implicit Euler to Semi-Explicit Model. Another possibility is to apply the linearly-implicit Euler discretization (5.2) to the differential-algebraic system (6.16). This gives

$$\begin{pmatrix} I & -hI \\ -hJ & hC(y_0) \end{pmatrix} \begin{pmatrix} y_{i+1} - y_i \\ z_{i+1} - z_i \end{pmatrix} = h \begin{pmatrix} f(y_i) - C(y_i)z_i \\ z_i \end{pmatrix} \quad (6.22)$$

with $z_0 = y'_0 = y'(x_0)$. The first line yields $z_{i+1} = (y_{i+1} - y_i)/h$ and the second line becomes

$$(C(y_0) - hJ)(y_{i+1} - y_i) = hf(y_i) - (C(y_i) - C(y_0))(y_i - y_{i-1}). \quad (6.23)$$

The right-most term vanishes for $i = 0$, so that y_{-1} does not enter the algorithm.

Asymptotic Expansions. The theoretical justification of the use of either (6.20) or (6.23) as basic step for an extrapolation process requires the investigation of the asymptotic expansion of their global errors.

In the situation where $C(y)$ is invertible, the discretization (6.20) is a consistent one-step discretization of (6.1') and possesses therefore, by standard theory (Theorem II.8.1), an asymptotic expansion, the terms of which, however, depend on the stiffness. Since the system (6.16) is of the form (1.6) with assumption (1.7) satisfied, we can conclude from Theorem 5.3 the existence of a *perturbed* asymptotic expansion for the second discretization (6.23).

In the situation where $C(y)$ is singular, Lubich (1989) revealed the existence of a perturbed asymptotic expansion for both discretizations (6.20) and (6.23). We refer to this original publication for further details, in particular to the study of the influence of these perturbations to the extrapolated numerical approximation.

Exercises

1. Reconstruct E. Hopf's analytic solution of Burgers' equation (6.6).

Hint. Introduce the new dependent variable

$$\varphi(x, t) = \exp \left\{ -\frac{1}{2\mu} \int_0^x u(\xi, t) d\xi - \int_0^t c(\tau) d\tau \right\}.$$

Show that for a suitably chosen $c(t)$ the function $\varphi(x, t)$ satisfies the one dimensional heat equation. The solution $u(x, t)$ of (6.6) can then be recovered from $\varphi(x, t)$ by

$$u = -2\mu(\log \varphi)_x = -2\mu(\varphi_x/\varphi).$$

2. Assume that (6.8) and (6.13) hold. By eliminating from $0 = C(y)z - f(y)$ as many components of z as possible, transform the system (6.16) into an equivalent one of the form

$$y' = F(y, u), \quad 0 = G(y),$$

where u collects the remaining components of z .

- a) Prove that Runge-Kutta methods and multistep methods are invariant with respect to this transformation.
 b) Show that $G_y(y)F_u(y, u)$ is invertible, so that the convergence results of Sections VII.3 and VII.4 can be applied.
3. (Quasilinear problems with gradient-type mass-matrix, see HLR89, page 111). Consider the electrical circuit (1.14), but suppose this time that the capacities depend on the voltages, e.g., as

$$C_k = C_{k0}/(1 - (U_i - U_j)/U_b)^{1/2}$$

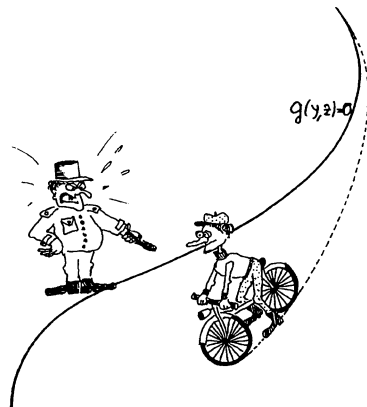
so that the expressions $C_k(U'_i - U'_j)$ in (1.14) must be replaced by $(C_k(U_i - U_j))'$. Show that then the corresponding equations are of the form (6.1) with

$$C(y) = Aq'(y)$$

where A is a constant matrix and $q(y)$ a known function of y . Show that such problems can be efficiently solved by introducing $q(y) = z$ as a new variable such that the problem becomes semi-explicit as

$$Az' = f(y), \quad 0 = z - q(y).$$

Chapter VII. Differential-Algebraic Equations of Higher Index



(Drawing by K. Wanner)

In the preceding chapter we considered the simplest special case of differential-algebraic equations – the so-called index 1 problem. Many problems of practical interest are, however, of higher index, which makes them more and more difficult for their numerical treatment.

We start by classifying differential-algebraic equations (DAE) by the index (index of nilpotency for linear problems with constant coefficients; differentiation and perturbation index for general nonlinear problems) and present some examples arising in applications (Sect. VII.1). Several different approaches for solving numerically higher index problems are discussed in Sect. VII.2: index reduction by differentiation combined with suitable projections, state space form methods, and treatment as overdetermined or unstructured systems. Sections VII.3 and VII.4 study the convergence properties of multistep methods and Runge-Kutta methods when they are applied directly to index 2 systems. It may happen that the order of convergence is lower than for ordinary differential equations (“order reduction”). The study of conditions which guarantee a certain order is the subject of Sect. VII.5. Half-explicit methods for index 2 problems are especially suited for constrained mechanical systems (Sect. VII.6). A multibody mechanism and its numerical treatment are detailed in Sect. VII.7. Finally, we discuss symplectic methods for constrained Hamiltonian systems (Sect. VII.8), and explain their long-term behaviour by a backward error analysis for differential equations on manifolds.

VII.1 The Index and Various Examples

The most general form of a differential-algebraic system is that of an implicit differential equation

$$F(u', u) = 0 \quad (1.1)$$

where F and u have the same dimension. We always assume F to be sufficiently differentiable. A non-autonomous system is brought to the form (1.1) by appending x to the vector u , and by adding the equation $x' = 1$.

If $\partial F / \partial u'$ is invertible we can formally solve (1.1) for u' to obtain an ordinary differential equation. In this chapter we are interested in problems (1.1) where $\partial F / \partial u'$ is singular.

Linear Equations with Constant Coefficients

Uebrigens kann ich die Meinung des Hrn. *Jordan* nicht theilen, dass es ziemlich schwer sei, der *Weierstrass*-schen Analyse zu folgen; sie scheint mir im Gegentheil vollkommen durchsichtig zu sein, ...
(L. Kronecker 1874)

The simplest and best understood problems of the form (1.1) are linear differential equations with constant coefficients

$$Bu' + Au = d(x). \quad (1.2)$$

In looking for solutions of the form $e^{\lambda x} u_0$ (if $d(x) \equiv 0$) we are led to consider the “matrix pencil” $A + \lambda B$. When $A + \lambda B$ is singular for all values of λ , then (1.2) has either no solution or infinitely many solutions for a given initial value (Exercise 1). We shall therefore deal only with *regular matrix pencils*, i.e., with problems where the polynomial $\det(A + \lambda B)$ does not vanish identically. The key to the solution of (1.2) is the following simultaneous transformation of A and B to canonical form.

Theorem 1.1 (Weierstrass 1868, Kronecker 1890). *Let $A + \lambda B$ be a regular matrix pencil. Then there exist nonsingular matrices P and Q such that*

$$PAQ = \begin{pmatrix} C & 0 \\ 0 & I \end{pmatrix}, \quad PBQ = \begin{pmatrix} I & 0 \\ 0 & N \end{pmatrix} \quad (1.3)$$

where $N = \text{blockdiag}(N_1, \dots, N_k)$, each N_i is of the form

$$N_i = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & 0 & 1 \\ 0 & & & 0 \end{pmatrix}, \quad \text{of dimension } m_i, \quad (1.4)$$

and C can be assumed to be in Jordan canonical form.

Proof (Gantmacher 1954 (Chapter XII), see also Exercises 2 and 3). We fix some c such that $A + cB$ is invertible. If we multiply

$$A + \lambda B = A + cB + (\lambda - c)B$$

by the inverse of $A + cB$ and then transform $(A + cB)^{-1}B$ to Jordan canonical form (Theorem I.12.2) we obtain

$$\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + (\lambda - c) \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix}. \quad (1.5)$$

Here, J_1 contains the Jordan blocks with non-zero eigenvalues, J_2 those with zero eigenvalues (the dimension of J_1 is just the degree of the polynomial $\det(A + \lambda B)$). Consequently, J_1 and $I - cJ_2$ are both invertible and multiplying (1.5) from the left by $\text{blockdiag}(J_1^{-1}, (I - cJ_2)^{-1})$ gives

$$\begin{pmatrix} J_1^{-1}(I - cJ_1) & 0 \\ 0 & I \end{pmatrix} + \lambda \begin{pmatrix} I & 0 \\ 0 & (I - cJ_2)^{-1}J_2 \end{pmatrix}.$$

The matrices $J_1^{-1}(I - cJ_1)$ and $(I - cJ_2)^{-1}J_2$ can then be brought to Jordan canonical form. Since all eigenvalues of $(I - cJ_2)^{-1}J_2$ are zero, we obtain the desired decomposition (1.3). \square

Theorem 1.1 allows us to solve (1.2) as follows: we premultiply (1.2) by P and use the transformation

$$u = Q \begin{pmatrix} y \\ z \end{pmatrix}, \quad Pd(x) = \begin{pmatrix} \eta(x) \\ \delta(x) \end{pmatrix}.$$

This decouples the differential-algebraic system (1.2) into

$$y' + Cy = \eta(x), \quad Nz' + z = \delta(x). \quad (1.6)$$

The equation for y is just an ordinary differential equation. The relation for z decouples again into k subsystems, each of the form (with $m = m_i$)

$$\begin{aligned} z'_2 + z_1 &= \delta_1(x) \\ &\vdots \\ z'_m + z_{m-1} &= \delta_{m-1}(x) \\ z_m &= \delta_m(x). \end{aligned} \quad (1.7)$$

Here z_m is determined by the last equation, and the other components are obtained recursively by repeated differentiation. Thus z_1 depends on the $(m-1)$ -th derivative of $\delta_m(x)$. Since numerical differentiation is an unstable procedure, the largest m_i appearing in (1.4) is a measure of numerical difficulty for solving problem (1.2). This integer ($\max m_i$) is called the *index of nilpotency* of the matrix pencil $A + \lambda B$. It does not depend on the particular transformation used to get (1.3) (see Exercise 4).

Linear Equations with Variable Coefficients. In the case, where the matrices A and B in (1.2) depend on x , the study of the solutions is much more complicated. Multiplying the equation by $P(x)$ and substituting $u = Q(x)v$, yields the system

$$PBQv' + (PAQ + PBQ')v = 0, \quad (1.8)$$

which shows that the transformation (1.3) is no longer relevant. With the use of transformations of the form (1.8), Kunkel & Mehrmann (1995) derive a canonical form for linear systems with variable coefficients.

Differentiation Index

A lot of English cars have steering wheels.

(*Fawlty Towers*, Cleese and Booth 1979)

Let us start with the following example:

$$\begin{aligned} y_1' &= 0.7 \cdot y_2 + \sin(2.5 \cdot z) = f_1(y, z) \\ y_2' &= 1.4 \cdot y_1 + \cos(2.5 \cdot z) = f_2(y, z) \end{aligned} \quad (1.9a)$$

$$0 = y_1^2 + y_2^2 - 1 = g(y). \quad (1.9b)$$

The “control variable” z in (1.9a) can be interpreted as the position of a “steering wheel” keeping the vector field (y_1', y_2') tangent to the circle $y_1^2 + y_2^2 = 1$, so that condition (1.9b) remains continually satisfied (see Fig. 1.1a). By differentiating (1.9b) and substituting (1.9a) we therefore must have

$$g_y(y)f(y, z) = 0. \quad (1.9c)$$

This defines a “hidden” submanifold of the cylinder, on which all solutions of (1.9a,b) must lie (see Fig. 1.1b). We still do not know how, with increasing x , the variable z changes. This is obtained by differentiating (1.9c) with respect to x : $g_{yy}(f, f) + g_y f_y f + g_y f_z z' = 0$. From this relation we can extract

$$z' = -(g_y f_z)^{-1} (g_{yy}(f, f) + g_y f_y f) \quad (1.9d)$$

if

$$g_y(y)f_z(y, z) \quad \text{is invertible.} \quad (1.10)$$

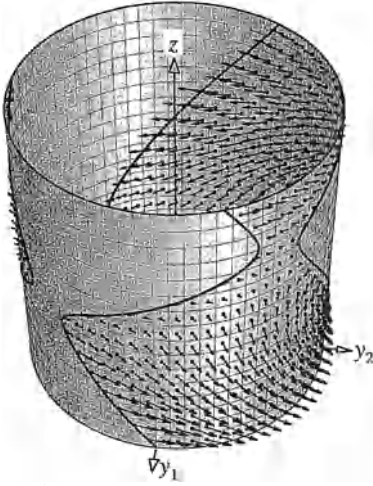


Fig. 1.1a. The vector field (1.9a,d)

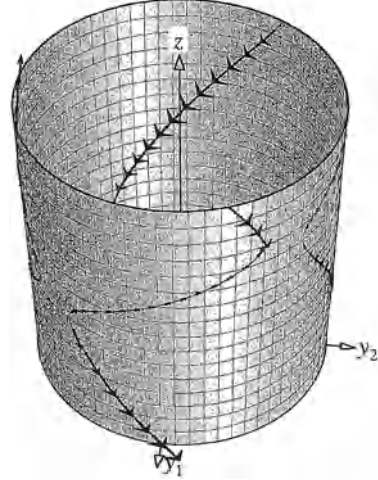


Fig. 1.1b. The hidden submanifold

We have been able to transform the above differential-algebraic equation (1.9a,b) into an ordinary differential system (1.9a,d) by *two analytic differentiations* of the constraint (1.9c). This fact is used for the following definition, which has been developed in several papers (Gear & Petzold 1983, 1984; Gear, Gupta & Leimkuhler 1985, Gear 1990, Campbell & Gear 1995).

Definition 1.2. Equation (1.1) has *differentiation index* $di = m$ if m is the minimal number of analytical differentiations

$$F(u', u) = 0, \quad \frac{dF(u', u)}{dx} = 0, \quad \dots, \quad \frac{d^m F(u', u)}{dx^m} = 0 \quad (1.11)$$

such that equations (1.11) allow us to extract by algebraic manipulations an explicit ordinary differential system $u' = \varphi(u)$ (which is called the “*underlying ODE*”).

Examples. Linear Equations with Constant Coefficients. The following problem

$$\begin{aligned} z_2' + z_1 &= \delta_1 & z_2'' + z_1' &= \delta_1' \\ z_3' + z_2 &= \delta_2 & z_3''' + z_2'' &= \delta_2'' & \Rightarrow & z_1' = \delta_1' - \delta_2'' + \delta_3''' \\ z_3 &= \delta_3 & z_3''' &= \delta_3''' \end{aligned} \quad (1.12)$$

can be seen to have differentiation index 3. For linear equations with constant coefficients the differentiation index and the index of nilpotency are therefore the same.

Systems of Index 1. The differential-algebraic systems already seen in Chapter VI

$$y' = f(y, z) \quad (1.13a)$$

$$0 = g(y, z) \quad (1.13b)$$

have no z' . We therefore differentiate (1.13b) to obtain

$$z' = -g_z^{-1}(y, z)g_y(y, z)f(y, z) \quad (1.13c)$$

which is possible if g_z is invertible in a neighbourhood of the solution. The problem (1.13a,b), for invertible g_z , is thus of differentiation index 1.

Systems of Index 2. In the system (see example (1.9))

$$y' = f(y, z) \quad (1.14a)$$

$$0 = g(y), \quad (1.14b)$$

where the variable z is absent in the algebraic constraint, we obtain by differentiation of (1.14b) the "hidden constraint"

$$0 = g_y(y)f(y, z). \quad (1.14c)$$

If (1.10) is satisfied in a neighbourhood of the solution, then (1.14a) and (1.14c) constitute an index 1 problem. Differentiation of (1.14c) yields the missing differential equation for z , so that the problem (1.14a,b) is of differentiation index 2. If the initial values satisfy $0 = g(y_0)$ and $0 = g_y(y_0)f(y_0, z_0)$, we call them *consistent*. In this case, and only in this case, the system (1.14a,b) possesses a (locally) unique solution.

System (1.14a,b) is a representative of the larger class of problems of type (1.13a,b) with *singular* g_z . If we assume that g_z has constant rank in a neighbourhood of the solution, we can eliminate certain algebraic variables from $0 = g(y, z)$ until the system is of the form (1.14). This can be done as follows: from the constant rank assumption it follows that either there exists a component of g such that $\partial g_i / \partial z_1 \neq 0$ locally, or $\partial g / \partial z_1$ vanishes identically so that g is already independent of z_1 . In the first case we can express z_1 as a function of y and the remaining components of z , and then we can eliminate z_1 from the system. Repeating this procedure with z_2, z_3 , etc., will lead to a system of the form (1.14). This transformation does not change the index. Moreover, most numerical methods are invariant under this transformation. Therefore, theoretical work done for systems of the form (1.14) will also be valid for more general problems.

Systems of Index 3. Problems of the form

$$y' = f(y, z) \quad (1.15a)$$

$$z' = k(y, z, u) \quad (1.15b)$$

$$0 = g(y) \quad (1.15c)$$

are of differentiation index 3, if

$$g_y f_z k_u \quad \text{is invertible} \quad (1.16)$$

in a neighbourhood of the solution. Differentiating (1.15c) twice gives

$$0 = g_y f \quad (1.15d)$$

$$0 = g_{yy}(f, f) + g_y f_y f + g_y f_z k. \quad (1.15e)$$

Equations (1.15a,b), (1.15e) together with Condition (1.16) are of the index 1 form (1.13a,b). Consistent initial values must satisfy the three conditions (1.15c,d,e).

An extensive study of the solution space of general differential-algebraic systems is done by Griepentrog & März (1986), März (1989, 1990). These authors try to avoid assumptions on the smoothness on the problem as far as possible and replace the above differentiations by a careful study of suitable projections depending only on the first derivatives of F .

Differential Equations on Manifolds

In the language of differentiable manifolds, whose use in DAE theory was urged by Rheinboldt (1984), a constraint (such as $g(y) = 0$) represents a manifold, which we denote by

$$\mathcal{M} = \{y \in \mathbb{R}^n \mid g(y) = 0\}. \quad (1.17)$$

We assume that $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ (with $m < n$) is a sufficiently differentiable function whose Jacobian $g_y(y)$ has full rank for $y \in \mathcal{M}$. For a fixed $y \in \mathcal{M}$ we denote by

$$T_y \mathcal{M} = \{v \in \mathbb{R}^n \mid g_y(y)v = 0\}, \quad (1.18)$$

the tangent space of \mathcal{M} at y . This is a linear space and has the same dimension $n - m$ as the manifold \mathcal{M} .

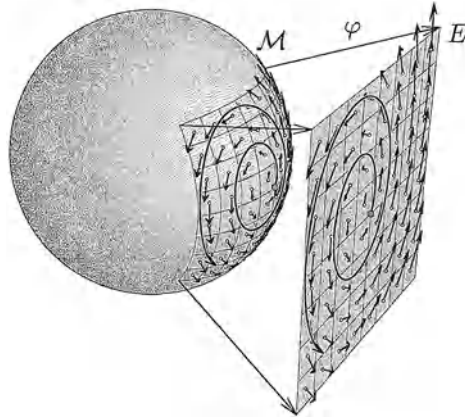


Fig. 1.2. A manifold with a tangent vector field, a chart, and a solution curve

A *vector field on \mathcal{M}* is a mapping $v: \mathcal{M} \rightarrow \mathbb{R}^n$, which satisfies $v(y) \in T_y \mathcal{M}$ for all $y \in \mathcal{M}$. For such a vector field we call

$$y' = v(y), \quad y \in \mathcal{M} \quad (1.19)$$

a *differential equation on the manifold \mathcal{M}* . Differentiation on an $(n-m)$ -dimensional manifold is described by so-called *charts* $\varphi_i: U_i \rightarrow E_i$, where the U_i cover

the manifold \mathcal{M} and the E_i are open subsets of \mathbb{R}^{n-m} (Fig. 1.2; see also Lang (1962), Chap. II and Abraham, Marsden & Ratiu (1983), Chap. III). The local theory of ordinary differential equations can be extended to vector fields on manifolds in a straightforward manner:

Project the vectors $v(y)$ onto E_i via a chart φ_i by multiplying $v(y)$ with the Jacobian of φ_i at y . Then apply standard results to the projected vector field in \mathbb{R}^{n-m} , and pull the solution back to \mathcal{M} .

(see Fig. 1.2). The local existence of solutions of (1.19) can be shown in this way. The obtained solution is independent of the chosen chart. Where the solution leaves the domain of a chart, the integration must be continued via another one.

Index 2 Problems. Consider the system (1.14a,b) and suppose that (1.10) is satisfied. This condition implies that $g_y(y)$ is of full rank, so that (1.17) is a smooth manifold. Moreover, the Implicit Function Theorem implies that the differentiated constraint (1.14c) can be solved for z (in a neighbourhood of the solution), i.e., there exists a smooth function $h(y)$ such that

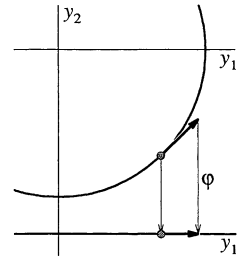
$$g_y(y)f(y, z) = 0 \quad \Longleftrightarrow \quad z = h(y). \quad (1.20)$$

Inserting this relation into (1.10a) yields

$$y' = f(y, h(y)), \quad y \in \mathcal{M} \quad (1.21)$$

which is a differential equation on the manifold (1.17), because $f(y, h(y)) \in T_y \mathcal{M}$ by (1.20). The differential equation (1.21) is equivalent to (1.14a,b).

Example. The manifold \mathcal{M} for problem (1.9) is one-dimensional (circle). In points, where $y_1 \neq \pm 1$, we can solve (1.9b) to obtain locally $y_2 = \pm \sqrt{1 - y_1^2}$. The map $(y_1, y_2) \mapsto y_1$ constitutes a chart φ , which is bijective in a neighbourhood of the considered point. Inserting z from (1.9c) and the above y_2 into (1.9a), yields an equation $y_1' = G(y_1)$, which is the projected vector field in \mathbb{R}^1 .



Index 3 Problems. For the system (1.15a,b,c) the solutions lie on the manifold

$$\mathcal{M} = \{(y, z) \mid g(y) = 0, \quad g_y(y)f(y, z) = 0\}. \quad (1.22)$$

The assumption (1.16) implies that $g_y(y)$ and $g_y(y)f_z(y, z)$ have full rank, so that \mathcal{M} is a manifold. Its tangent space at (y, z) is

$$T_{(y,z)} \mathcal{M} = \{(v, w) \mid g_y(y)v = 0, \quad g_{yy}(y)(f(y, z), v) + g_y(y)(f_y(y, z)v + f_z(y, z)w) = 0\}. \quad (1.23)$$

Solving Eq. (1.15e) for u and inserting the result into (1.15b) yields a differential equation on the manifold \mathcal{M} . Because of (1.15d,e), the obtained vector field lies in the tangent space $T_{(y,z)} \mathcal{M}$ for all $(y, z) \in \mathcal{M}$.

The Perturbation Index

Now fills thy sleep with perturbations.

(The *Ghost of Anne* in Shakespeare's *Richard III*, act V, sc. III)

A second concept of index, due to HLR89¹, interprets the index as a measure of sensitivity of the solutions with respect to perturbations of the given problem.

Definition 1.3. Equation (1.1) has *perturbation index* $pi = m$ along a solution $u(x)$ on $[0, \bar{x}]$, if m is the smallest integer such that, for all functions $\hat{u}(x)$ having a defect

$$F(\hat{u}', \hat{u}) = \delta(x), \quad (1.24)$$

there exists on $[0, \bar{x}]$ an estimate

$$\|\hat{u}(x) - u(x)\| \leq C \left(\|\hat{u}(0) - u(0)\| + \max_{0 \leq \xi \leq x} \|\delta(\xi)\| + \dots + \max_{0 \leq \xi \leq x} \|\delta^{(m-1)}(\xi)\| \right) \quad (1.25)$$

whenever the expression on the right-hand side is sufficiently small.

Remark. We deliberately do not write “Let $\hat{u}(x)$ be the solution of $F(\hat{u}', \hat{u}) = \delta(x) \dots$ ” in this definition, because the existence of such a solution $\hat{u}(x)$ for an arbitrarily given $\delta(x)$ is not assured. We start with \hat{u} and then compute δ as defect of (1.1).

Systems of Index 1. For the computation of the perturbation index of (1.13a,b) we consider the perturbed system

$$\hat{y}' = f(\hat{y}, \hat{z}) + \delta_1(x) \quad (1.26a)$$

$$0 = g(\hat{y}, \hat{z}) + \delta_2(x). \quad (1.26b)$$

The essential observation is that the difference $\hat{z} - z$ can be estimated with the help of the Implicit Function Theorem, without any differentiation of the equation. Since g_z is invertible by hypothesis, this theorem gives from (1.26b) compared to (1.13b)

$$\|\hat{z}(x) - z(x)\| \leq C_1 (\|\hat{y}(x) - y(x)\| + \|\delta_2(x)\|) \quad (1.27)$$

as long as the right-hand side of (1.27) is sufficiently small. We now subtract (1.26a) from (1.13a), integrate from 0 to x , use a Lipschitz condition for f and the above estimate for $\hat{z}(x) - z(x)$. This gives for $e(x) = \|\hat{y}(x) - y(x)\|$:

$$e(x) \leq e(0) + C_2 \int_0^x e(t) dt + C_3 \int_0^x \|\delta_2(t)\| dt + \left\| \int_0^x \delta_1(t) dt \right\|.$$

In this estimate the norm is *inside* the integral for δ_2 , but *outside* the integral for δ_1 . This is due to the fact that perturbations of the algebraic equation (1.13b) are more

¹ The “Lecture Notes” of Hairer, Lubich & Roche (1989) will be cited frequently in the subsequent sections. Reference to this publication will henceforth be denoted by HLR89.

serious than perturbations of the differential equation (1.13a). We finally apply Gronwall's Lemma (Exercise I.10.2) to obtain on a bounded interval $[0, \bar{x}]$

$$\begin{aligned}\|\widehat{y}(x) - y(x)\| &\leq C_4 \left(\|\widehat{y}(0) - y(0)\| + \int_0^x \|\delta_2(t)\| dt + \max_{0 \leq \xi \leq x} \left\| \int_0^\xi \delta_1(t) dt \right\| \right) \\ &\leq C_5 \left(\|\widehat{y}(0) - y(0)\| + \max_{0 \leq \xi \leq x} \|\delta_2(\xi)\| + \max_{0 \leq \xi \leq x} \|\delta_1(\xi)\| \right).\end{aligned}$$

This inequality, together with (1.27), shows that the perturbation index of the problem is 1.

Systems of Index 2. We consider the following perturbation of system (1.14a,b)

$$\widehat{y}' = f(\widehat{y}, \widehat{z}) + \delta(x) \quad (1.28a)$$

$$0 = g(\widehat{y}) + \theta(x). \quad (1.28b)$$

Differentiation of (1.28b) gives

$$0 = g_y(\widehat{y})f(\widehat{y}, \widehat{z}) + g_y(\widehat{y})\delta(x) + \theta'(x). \quad (1.29)$$

Under the assumption (1.10) we can use the estimates of the index 1 case (with $\delta_2(x)$ replaced by $g_y(\widehat{y}(x))\delta(x) + \theta'(x)$) to obtain

$$\begin{aligned}\|\widehat{y}(x) - y(x)\| &\leq C \left(\|\widehat{y}(0) - y(0)\| + \int_0^x (\|\delta(\xi)\| + \|\theta'(\xi)\|) d\xi \right) \\ \|\widehat{z}(x) - z(x)\| &\leq C \left(\|\widehat{y}(0) - y(0)\| + \max_{0 \leq \xi \leq x} \|\delta(\xi)\| + \max_{0 \leq \xi \leq x} \|\theta'(\xi)\| \right).\end{aligned} \quad (1.30)$$

Since these estimates depend on the first derivative of θ , the perturbation index of this problem is 2. A sharper estimate for the y -component is given in Exercise 6.

Example. Fig. 1.3 presents an illustration for the index 2 problem (1.9a,b). Small perturbations of $g(y)$, once discontinuous in the first derivative (left), the other of oscillatory type (right), results in discontinuities or violent oscillations of z , respectively.

The above examples might give the impression that the differentiation index and the perturbation index are always equal. The following counter-examples show that this is not true.

Counterexamples. The first counterexample of type $M(y)y' = f(y)$ is given by Lubich (1989):

$$\begin{aligned}y_1' - y_3 y_2' + y_2 y_3' &= 0 & \widehat{y}_1' - \widehat{y}_3 \widehat{y}_2' + \widehat{y}_2 \widehat{y}_3' &= 0 \\ y_2 &= 0 & \widehat{y}_2 &= \varepsilon \sin \omega x \\ y_3 &= 0 & \widehat{y}_3 &= \varepsilon \cos \omega x\end{aligned} \quad (1.31)$$

with $y_i(0) = 0$ ($i = 1, 2, 3$). Inserting $\widehat{y}_2 = \varepsilon \sin \omega x$ and $\widehat{y}_3 = \varepsilon \cos \omega x$ into the first equation gives $\widehat{y}_1' = \varepsilon^2 \omega$ which makes, for ε fixed and $\omega \rightarrow \infty$, an estimate

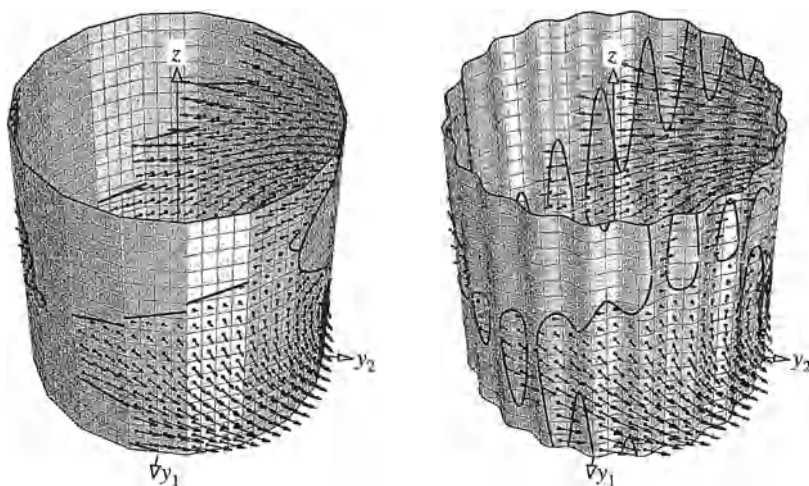


Fig. 1.3. Perturbations of an index 2 problem

(1.25) with $m = 1$ impossible. However, for $m = 2$ the estimate (1.25) is clearly satisfied. This problem, which is obviously of differentiation index 1, is thus of perturbation index 2.

It was believed for some time (see the first edition, p. 479), that the differentiation and perturbation indices can differ at most by 1. The following example, due to Campbell & Gear (1995), was therefore a big surprise:

$$y_m N y' + y = 0, \quad (1.32)$$

where N is a $m \times m$ upper triangular nilpotent Jordan block. Since the last row of N is zero, we have $y_m = 0$, and the differentiation index is 1. On the other hand, adding a perturbation makes y_m different from zero. This is the reason why the perturbation index of (1.32) is m .

Control Problems

Many problems of control theory lead to ordinary differential equations of the form $y' = f(y, u)$, where u represents a set of controls. Similar as in example (1.9) above, these controls must be applied so that the solution satisfies some constraints $0 = g(y, u)$. For numerical examples of such control problems we refer to Brenan (1983) (space shuttle simulation) and Brenan, Campbell & Petzold (1989).

Optimal Control Problems are differential equations $y' = f(y, u)$ formulated in such a way that the control $u(x)$ has to minimize some cost functional. The Euler-Lagrange equation then often becomes a differential-algebraic system (Pontryagin, Boltyanskij, Gamkrelidze & Mishchenko 1961, Athans & Falb 1966, Campbell 1982). We demonstrate this on the problem

$$y' = f(y, u), \quad y(0) = y_0 \quad (1.33a)$$

with cost functional

$$J(u) = \int_0^1 \varphi(y(x), u(x)) dx. \quad (1.33b)$$

For a given function $u(x)$ the solution $y(x)$ is determined by (1.33a). In order to find conditions for $u(x)$ which minimize $J(u)$ of (1.33b), we consider the perturbed control $u(x) + \varepsilon \delta u(x)$ where $\delta u(x)$ is an arbitrary function and ε a small number. To this control there corresponds a solution $y(x) + \varepsilon \delta y(x) + \mathcal{O}(\varepsilon^2)$ of (1.33a); hence (by comparing powers of ε)

$$\delta y'(x) = f_y(x) \delta y(x) + f_u(x) \delta u(x), \quad \delta y(0) = 0, \quad (1.34)$$

where, as usual, $f_y(x) = f_y(y(x), u(x))$, etc. Linearization of (1.33b) shows that

$$J(u + \varepsilon \delta u) - J(u) = \varepsilon \int_0^1 (\varphi_y(x) \delta y(x) + \varphi_u(x) \delta u(x)) dx + \mathcal{O}(\varepsilon^2)$$

so that

$$\int_0^1 (\varphi_y(x) \delta y(x) + \varphi_u(x) \delta u(x)) dx = 0 \quad (1.35)$$

is a necessary condition for $u(x)$ to be an optimal solution of our problem. In order to express δy in terms of δu in (1.35), we introduce the adjoint differential equation

$$v' = -f_y(x)^T v - \varphi_y(x)^T, \quad v(1) = 0 \quad (1.36)$$

with inhomogeneity $\varphi_y(x)^T$. Hence we have (see Exercise 7)

$$\int_0^1 \varphi_y(x) \delta y(x) dx = \int_0^1 v^T(x) f_u(x) \delta u(x) dx. \quad (1.37)$$

Inserted into (1.35) this gives the necessary condition

$$\int_0^1 (v^T(x) f_u(x) + \varphi_u(x)) \delta u(x) dx = 0. \quad (1.38)$$

Since this relation has to be satisfied for all δu we obtain the necessary relation $v^T(x) f_u(x) + \varphi_u(x) = 0$ by the so-called “fundamental lemma of variational calculus”.

In summary, we have proved that a solution of the above optimal control problem has to satisfy the system

$$\begin{aligned} y' &= f(y, u), & y(0) &= y_0 \\ v' &= -f_y(y, u)^T v - \varphi_y(y, u)^T, & v(1) &= 0 \\ 0 &= v^T f_u(y, u) + \varphi_u(y, u). \end{aligned} \quad (1.39)$$

This is a boundary value differential-algebraic problem. It can also be obtained directly from the Pontryagin minimum principle (see Pontryagin et al. 1961, Athans & Falb 1966).

Differentiation of the algebraic relation in (1.39) shows that the system (1.39) has index 1 if the matrix

$$\sum_{i=1}^n v_i \frac{\partial^2 f_i}{\partial u^2}(y, u) + \frac{\partial^2 \varphi}{\partial u^2}(y, u) \quad (1.40)$$

is invertible along the solution. A situation where the system (1.39) has index 3 is presented in Exercise 8. An index 5 problem of this type is given in “Example 3.1” of Clark (1988). Other control problems with a large index are discussed in Campbell (1995).

Mechanical Systems

... berechnen wir T, V, L . Mehr brauchen wir von der Geometrie und Mechanik unseres Systems nicht zu wissen. Alles übrige besorgt ohne unser Zutun der Formalismus von LAGRANGE.
(Sommerfeld 1942, §35)

An interesting class of differential-algebraic systems appears in mechanical modeling of constrained systems. A choice method for deriving the equations of motion of mechanical systems is the Lagrange-Hamilton principle, whose long history goes back to merely theological ideas of Leibniz and Maupertuis. Let q_1, \dots, q_n be position coordinates of a system and $u_i = \dot{q}_i$ the velocities. Suppose a function $L(q, \dot{q})$ is given; then the Euler equations of the variational problem

$$\int_{t_1}^{t_2} L(q, \dot{q}) dt = \min ! \quad (1.41)$$

are given by

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_k} \right) - \frac{\partial L}{\partial q_k} = 0, \quad k = 1, \dots, n \quad (1.42)$$

or

$$\sum_{\ell=1}^n L_{\dot{q}_k \dot{q}_\ell} \ddot{q}_\ell = L_{q_k} - \sum_{\ell=1}^n L_{\dot{q}_k q_\ell} \dot{q}_\ell. \quad (1.43)$$

The great discovery of Lagrange (1788) is that for $L = T - U$, where T is the *kinetic energy* and U the *potential energy*, the differential equations (1.43) describe the movement of the corresponding “conservative system”. For a proof and various generalizations, consult any book on mechanics e.g., Sommerfeld (1942), vol. I, §§ 33–37, or Arnol’d (1979), part II.

Example 1. For the mathematical pendulum of length ℓ we choose as position coordinate the angle $\theta = q_1$ such that $T = m\ell^2 \dot{\theta}^2 / 2$ and $U = -\ell mg \cos \theta$. Then (1.43) becomes $\ell \ddot{\theta} = -g \sin \theta$, the well-known pendulum equation.

Movement with Constraints. Suppose now that we have some constraints $g_1(q) = 0, \dots, g_m(q) = 0$ on our movement. Another great idea of Lagrange is to vary the “Lagrange function” as follows in this case

$$L = T - U - \lambda_1 g_1(q) - \dots - \lambda_m g_m(q) \quad (1.44)$$

where the “Lagrange multipliers” λ_i are appended to the coordinates. The important fact is that, since L is independent of λ_i , the equation (1.43), for the derivatives with respect to λ_k , just becomes $0 = g_k(q)$, the desired side conditions.

Example 2. We now describe the pendulum in Cartesian coordinates x, y with constraint $x^2 + y^2 - \ell^2 = 0$. This gives for (1.44)

$$L = \frac{m}{2}(\dot{x}^2 + \dot{y}^2) - mgy - \lambda(x^2 + y^2 - \ell^2)$$

and (1.43) becomes

$$\begin{aligned} m\ddot{x} &= -2x\lambda \\ m\ddot{y} &= -mg - 2y\lambda \\ 0 &= x^2 + y^2 - \ell^2. \end{aligned} \quad (1.45)$$

In this example the physical meaning of λ is the tension in the rod which maintains the mass point on the desired orbit.

The general form of a constrained mechanical system (1.43) is in vector notation (after replacing dots by primes)

$$q' = u \quad (1.46a)$$

$$M(q)u' = f(q, u) - G^T(q)\lambda \quad (1.46b)$$

$$0 = g(q) \quad (1.46c)$$

where $M(q) = T_{\dot{q}\dot{q}} = T_{uu}$ is a positive definite matrix, $G(q) = \partial g / \partial q$ and $q = (q_1, \dots, q_n)^T$, $u = (\dot{q}_1, \dots, \dot{q}_n)^T$, $\lambda = (\lambda_1, \dots, \lambda_m)^T$. Various formulations are possible for such a problem, each of which leads to a different numerical approach.

Index 3 Formulation (position level, descriptor form). If we formally multiply (1.46b) by M^{-1} , the system (1.46) becomes of the form (1.15) with (q, u, λ) in the roles of (y, z, u) . The condition (1.16), written out for (1.46), is

$$GM^{-1}G^T \quad \text{is invertible.} \quad (1.47)$$

This is satisfied, if the constraints (1.46c) are independent, i.e., if the rows of the matrix G are linearly independent. Under this assumption, the system (1.46a,b,c) is thus an index 3 problem.

Index 2 Formulation (velocity level). Differentiation of (1.46c) gives

$$0 = G(q)u. \quad (1.46d)$$

If we replace (1.46c) by (1.46d) we obtain a system of the form (1.14a,b) with (q, u) in the role of y and λ that of z . One verifies that Condition (1.10) is equivalent to (1.47), so that (1.46a,b,d) represents a problem of index 2.

Index 1 Formulation (acceleration level). If we differentiate twice the constraint (1.46c), the resulting equation together with (1.46b) yield

$$\begin{pmatrix} M(q) & G^T(q) \\ G(q) & 0 \end{pmatrix} \begin{pmatrix} u' \\ \lambda \end{pmatrix} = \begin{pmatrix} f(q, u) \\ -g_{qq}(q)(u, u) \end{pmatrix}. \quad (1.46e)$$

This allows us to express u' and λ as functions of q, u , provided that the matrix in (1.46e) is invertible. Hence, (1.46a,e) constitute an index 1 problem. The assumption on the matrix in Eq. (1.46e) is weaker than (1.47), because $M(q)$ need not be regular.

All these formulations are mathematically equivalent, if the initial values are consistent, i.e., if (1.46c,d,e) are satisfied. However, if for example the index 2 system (1.46a,b,d) is integrated numerically, the constraints of the original problem will no longer be exactly satisfied. For this reason Gear, Gupta & Leimkuhler (1985) introduced another index 2 formulation (“... an interesting way of reducing the problem to index two and adding variables so that the constraint continues to be satisfied”).

GGL Formulation. The idea is to add the constraint (1.46d) to the original system and to introduce an additional Lagrange multiplier μ in (1.46a). For the sake of symmetry we also multiply (1.46a) by $M(q)$, so that the whole system becomes

$$\begin{aligned} M(q)q' &= M(q)u - G^T(q)\mu \\ M(q)u' &= f(q, u) - G^T(q)\lambda \\ 0 &= g(q) \\ 0 &= G(q)u. \end{aligned} \quad (1.48)$$

Here the differential variables are (q, u) and the algebraic variables are (μ, λ) . System (1.48) is of the form (1.14a,b) and the index 2 assumption is satisfied if (1.47) holds.

A concrete mechanical system is described in detail, together with numerical results for all the above formulations, in Sect. VII.7.

Exercises

1. Prove that the initial value problem

$$Bu' + Au = 0, \quad u(0) = u_0 \quad (1.49)$$

has a unique solution if and only if the matrix pencil $A + \lambda B$ is regular.

Hint for the “only if” part. If n is the dimension of u , choose arbitrarily $n+1$ distinct λ_i and vectors $v_i \neq 0$ satisfying $(A + \lambda_i B)v_i = 0$. Then take a linear combination, such that $\sum \alpha_i v_i = 0$, but $\sum \alpha_i e^{\lambda_i x} v_i \neq 0$.

2. (Stewart 1972). Let $A + \lambda B$ be a regular matrix pencil. Show that there exist unitary matrices Q and Z such that

$$QAZ = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad QBZ = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix} \quad (1.50)$$

are both triangular. Further, the diagonal elements of A_{22} and B_{11} are all 1, those of B_{22} are all 0.

Hint (Compare with the Schur decomposition of Theorem I.12.1). Let λ_1 be a zero of $\det(A + \lambda B)$ and $v_1 \neq 0$ be such that $(A + \lambda_1 B)v_1 = 0$. Verify that $Bv_1 \neq 0$ and that

$$AZ_1 = Q_1 \begin{pmatrix} -\lambda_1 & * \\ 0 & \tilde{A} \end{pmatrix}, \quad BZ_1 = Q_1 \begin{pmatrix} 1 & * \\ 0 & \tilde{B} \end{pmatrix}$$

where Q_1, Z_1 are unitary matrices whose first columns are Bv_1 and v_1 , respectively. The matrix pencil $\tilde{A} + \lambda \tilde{B}$ is again regular and this procedure can be continued until $\det(\tilde{A} + \lambda \tilde{B}) = \text{Const}$ which implies that $\det \tilde{B} = 0$. In this case we take a vector $v_2 \neq 0$ such that $\tilde{B}v_2 = 0$ and transform $\tilde{A} + \lambda \tilde{B}$ with unitary matrices Q_2, Z_2 , whose first columns are $\tilde{A}v_2$ and v_2 , respectively. For a practical computation of the decomposition (1.50) see Golub & Van Loan (1989), Sect. 7.7.

3. Under the assumptions of Exercise 2 show that there exist matrices S and T such that

$$\begin{pmatrix} I & S \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} \begin{pmatrix} I & T \\ 0 & I \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix},$$

$$\begin{pmatrix} I & S \\ 0 & I \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix} \begin{pmatrix} I & T \\ 0 & I \end{pmatrix} = \begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix}.$$

Hint. These matrices have to satisfy

$$A_{11}T + A_{12} + SA_{22} = 0 \quad (1.51a)$$

$$B_{11}T + B_{12} + SB_{22} = 0 \quad (1.51b)$$

and can be computed as follows: the first column of T is obtained from (1.51b) because B_{11} is invertible and the first column of SB_{22} vanishes; then the first column of S is given by (1.51a) because A_{22} is invertible; the second column of SB_{22} is then known and we can compute the second column of T from (1.51b), etc.

4. Prove that the index of nilpotency of a regular matrix pencil $A + \lambda B$ does not depend on the choice of P and Q in (1.3).

Hint. Consider two different decompositions of the form (1.3) and denote the matrices which appear by C_1, N_1 and C_2, N_2 , respectively. Show the existence of a regular matrix T such that $N_2 = T^{-1}N_1T$.

5. Prove that the system (VI.3.4a,b) has index 2 (it is of the form (1.14a,b) and satisfies (1.10)). The full system (VI.3.4) has perturbation index k .
6. (Arnold 1993). Consider the index 2 problem (1.14) and its perturbation (1.28). Prove that the difference $\Delta y(x) = \hat{y}(x) - y(x)$ satisfies

$$\|\Delta y(x)\| \leq C \left(\|\Delta y(0)\| + \max_{0 \leq \xi \leq x} \left(\left\| \int_0^\xi P(t) \delta(t) dt \right\| + \|\theta(\xi)\| + (\|\delta(\xi)\| + \|\theta'(x)\|)^2 \right) \right)$$

with the projector $P(t) = I - (f_z(g_y f_z)^{-1} g_y)(y(t), z(t))$, provided that the right hand side is sufficiently small.

Hint. Linearize Eq. (1.29) around (y, z) , extract $\hat{z} - z$, and insert it into the difference of (1.28a) and (1.14a). The term $(f_z(g_y f_z)^{-1})(y(x), z(x)) \theta'(x)$ can be replaced by $\frac{d}{dx}(f_z(g_y f_z)^{-1}(y(x), z(x)) \theta(x)) + \mathcal{O}(\|\theta(x)\|)$ before integration.

7. For the linear initial value problem

$$y' = A(x)y + f(x), \quad y(0) = 0$$

consider the *adjoint* problem

$$v' = -A(x)^T v - g(x), \quad v(1) = 0.$$

Prove that
$$\int_0^1 g(x)^T y(x) dx = \int_0^1 v(x)^T f(x) dx.$$

8. Consider a linear optimal control problem with quadratic cost functional

$$\begin{aligned} y' &= Ay + Bu + f(x), & y(0) &= y_0 \\ J(u) &= \frac{1}{2} \int_0^1 \left(y(x)^T C y(x) + u(x)^T D u(x) \right) dx, \end{aligned}$$

where C and D are assumed to be positive semi-definite.

- a) Prove that $J(u)$ is minimal if and only if

$$\begin{aligned} y' &= Ay + Bu + f(x), & y(0) &= y_0 \\ v' &= -A^T v - Cy, & v(1) &= 0 \\ 0 &= B^T v + Du. \end{aligned} \tag{1.52}$$

- b) If D is positive definite, then (1.52) has index 1.

- c) If $D = 0$ and $B^T C B$ is positive definite, then (1.52) has index 3.

VII.2 Index Reduction Methods

We have seen in Sect. VI.1 that the numerical treatment of problems of index 1, which are either in the half-explicit form (1.13) or in the form $Mu' = \varphi(u)$, is not much more difficult than that of ordinary differential equations. For higher index problems the situation changes completely. This section is devoted to the study of several approaches that are all based on the idea of modifying the problem in such a way that the index is reduced.

Index Reduction by Differentiation

The most apparent way of reducing the index is to differentiate repeatedly the algebraic constraints (see Definition 1.2). In general, it is recommended to differentiate until having obtained an index 1 problem. For example, the index 2 problem (1.14a,b) is replaced by (1.14a,c), or the constrained mechanical system (1.46a,b,c) by (1.46a,b,e). The resulting problem is then solved by the methods of Chapter VI.

We illustrate this approach at the “pendulum example”

$$x' = u, \quad u' = -x\lambda \quad (2.1a)$$

$$y' = v, \quad v' = -1 - y\lambda \quad (2.1b)$$

$$0 = x^2 + y^2 - 1. \quad (2.1c)$$

In this form it has index 3. Differentiating the algebraic constraint twice yields

$$0 = xu + yv, \quad (2.2)$$

$$0 = -\lambda(x^2 + y^2) - y + u^2 + v^2. \quad (2.3)$$

Equations (2.1a,b) together with (2.3) represent an index 1 problem. We can extract λ from (2.3) and insert it into (2.1a,b) to get a differential equation for x, y, u, v , which can be solved by standard methods.

Drift-off Phenomenon. As an example we apply the code DOPRI5 to the index 1 problem (2.1a,b), (2.3) with initial values $x_0 = 1, y_0 = 0, u_0 = 0, v_0 = 0$. We are interested, how well the constraints (2.1c) and (2.2) are preserved by the numerical solution. The result presented in Fig. 2.1 shows that the error in the constraint (2.2) grows linearly, that in (2.1c) grows even quadratically. This phenomenon is explained as follows:

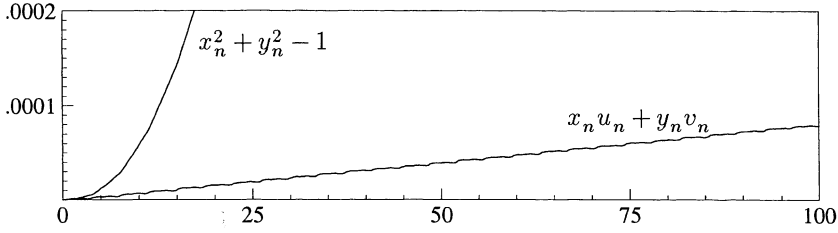


Fig. 2.1. Error in the constraints for DOPRI5 ($Atol = Rtol = 10^{-6}$)

Consider a constrained mechanical system (see (1.46))

$$q' = u \quad (2.4a)$$

$$M(q)u' = f(q, u) - G^T(q)\lambda \quad (2.4b)$$

$$0 = g(q). \quad (2.4c)$$

Differentiating (2.4c) twice we get

$$\begin{pmatrix} M(q) & G^T(q) \\ G(q) & 0 \end{pmatrix} \begin{pmatrix} u' \\ \lambda \end{pmatrix} = \begin{pmatrix} f(q, u) \\ -q_{qq}(q)(u, u) \end{pmatrix} \quad (2.5)$$

which, together with (2.4a), is the corresponding index 1 problem. The important observation is now that the index 1 problem possesses a solution for arbitrary initial values q_0 and u_0 . Due to the fact that the second derivative of $g(q(t))$ vanishes (this is a consequence of the lower relation of (2.5)), the solution of the index 1 problem satisfies

$$g(q(t)) = g(q_0) + (t - t_0)G(q_0)u_0, \quad (2.6a)$$

$$G(q(t))u(t) = G(q_0)u_0. \quad (2.6b)$$

Theorem 2.1. *If we apply a p th order numerical method to the index 1 problem (2.4a), (2.5) with consistent initial values at $t_0 = 0$, then the numerical solution (q_n, u_n) at time t_n satisfies (for $t_n - t_0 \leq \text{Const}$)*

$$\|g(q_n)\| \leq h^p(At_n + Bt_n^2), \quad \|G(q_n)u_n\| \leq h^p C t_n.$$

The value h represents the maximal step size used.

Proof. Denote by $q(t, t_0, q_0, u_0)$ the solution of the index 1 problem with initial value (q_0, u_0) at $t = t_0$. Since the local error $q_{j+1} - q(t_{j+1}, t_j, q_j, u_j)$ is of size $\mathcal{O}(h_j^{p+1})$ (and similarly for the u -component), it follows from (2.6a) that

$$\|g(q(t_n, t_{j+1}, q_{j+1}, u_{j+1})) - g(q(t_n, t_j, q_j, u_j))\| \leq h_j^{p+1}(A + 2B(t_n - t_{j+1})).$$

Adding up these inequalities from $j = 0$ to $j = n - 1$ gives the desired bound for $g(q_n)$, because the initial values are consistent, i.e., $g(q(t_n, t_0, q_0, u_0)) = 0$. The second estimate of Theorem 2.1 is proved in the same way. \square

Baumgarte Stabilization. The historically first remedy for this drift-off is due to Baumgarte (1972). Instead of replacing the constraint (2.4c) by its second time derivative, he proposes to replace (2.4c) by the linear combination

$$0 = \ddot{g} + 2\alpha\dot{g} + \beta^2 g, \quad (2.7)$$

where \dot{g} , \ddot{g} are time derivatives of (2.4c), i.e.,

$$g = g(q), \quad \dot{g} = G(q)u, \quad \ddot{g} = g_{qq}(q)(u, u) + G(q)(f(q, u) - G^T(q)\lambda).$$

Eq. (2.7) together with (2.4b) determines u' and λ as functions of (q, u) , and the resulting differential equation can be solved numerically. The idea is now to choose the free parameters α and β in such a way that (2.7) is an asymptotically stable differential equation, e.g., $\beta = \alpha$ and $\alpha > 0$. Consequently, the functions $g(q(t))$ and $G(q(t))u(t)$ are exponentially decreasing, in contrast to (2.6). The difficulty of this approach lies in a good choice of α . For small values of α the damping will not be sufficiently strong, whereas for large α the resulting differential equation becomes stiff and explicit methods are no longer efficient. A careful investigation on the choice of α can be found in Ascher, Chin & Reich (1994).

Stabilization by Projection

We shall now analyze another possibility for avoiding the instability of the preceding example, namely the repeated projection of the numerical solution onto the solution manifold.

Index 2 Problems. Consider the system (1.14a,b). Suppose that (y_{n-1}, z_{n-1}) is an approximation to the solution at time t_{n-1} which satisfies $g(y_{n-1}) = 0$ and $g_y(y_{n-1})f(y_{n-1}, z_{n-1}) = 0$. Applying a numerical one-step method (state space form method of Sect. VI.1) with these values to the index 1 system (1.14a,c) yields an approximation \tilde{y}_n, \tilde{z}_n that, in general, does not satisfy the constraint (1.14b). A natural way of projecting the approximation \tilde{y}_n to the solution manifold \mathcal{M} of Eq. (1.17) is along the image of f_z (see also the projected Runge-Kutta methods of Sect. VII.4). We therefore define y_n as the solution of

$$y - \tilde{y}_n = f_z(\tilde{y}_n, \tilde{z}_n)\mu, \quad g(y) = 0, \quad (2.8)$$

and then we adjust z_n by solving the equation $g_y(y_n)f(y_n, z_n) = 0$. Applying simplified Newton iterations to the nonlinear system (2.8) requires the decomposition of the matrix

$$\begin{pmatrix} I & f_z(\tilde{y}_n, \tilde{z}_n) \\ g_y(\tilde{y}_n) & 0 \end{pmatrix}. \quad (2.9)$$

Block elimination shows that the invertibility of (2.9) is a consequence of (1.10), and that only the matrix $g_y f_z$ has to be decomposed. Such a decomposition is usually already available from the application of the numerical method, so that the projection (2.8) is very cheap.

It is now natural to ask, whether this projection procedure can destroy the convergence properties of the underlying method. For a p th order one-step method the local error is of size $\mathcal{O}(h^{p+1})$. Since the solution of (1.14a,c) passing through (y_{n-1}, z_{n-1}) satisfies $g(y(t)) = 0$, it holds $g(\tilde{y}_n) = \mathcal{O}(h^{p+1})$. Hence, the solution of (2.8) satisfies $\mu = \mathcal{O}(h^{p+1})$, $y_n - \tilde{y}_n = \mathcal{O}(h^{p+1})$, and $z_n - \tilde{z}_n = \mathcal{O}(h^{p+1})$. By the Implicit Function Theorem this solution depends smoothly on $(\tilde{y}_n, \tilde{z}_n)$, so that the mapping $(y_{n-1}, z_{n-1}) \mapsto (y_n, z_n)$ represents a p th order one-step method for (1.14a,c). Convergence of order p thus follows from the standard theory (see Sects. VI.1 and II.3). This proof also applies to multistep methods.

Constrained Mechanical Systems. For the index 3 system (2.4a,b,c) the situation is slightly more complex. We assume consistent values $(q_{n-1}, u_{n-1}, \lambda_{n-1})$ at time t_{n-1} and apply a one-step method to the index 1 system (2.4a), (2.5) to obtain $(\tilde{q}_n, \tilde{u}_n)$. Since the position constraint (2.4c) only depends on q , the projections for q and u can be done sequentially.

Projection on Position Constraint. We define q_n as solution of the nonlinear system

$$\begin{aligned} M(\tilde{q}_n)(q_n - \tilde{q}_n) + G^T(\tilde{q}_n)\mu &= 0 \\ g(q_n) &= 0. \end{aligned} \quad (2.10)$$

Projection on Velocity Constraint. With the value q_n obtained from the above projection we let u_n be the solution of

$$\begin{aligned} M(q_n)(u_n - \tilde{u}_n) + G^T(q_n)\mu &= 0 \\ G(q_n)u_n &= 0. \end{aligned} \quad (2.11)$$

Lubich (1991) introduced this kind of projection, because “it is invariant under affine transformations of coordinates”. We remark that the system (2.11) is linear, whereas (2.10) is nonlinear and has to be solved by (simplified) Newton iterations. The index 3 assumption that the matrix in Eq. (2.5) is invertible, implies the existence of the projected values q_n and u_n (at least for sufficiently small step size). It is possible to alter slightly the arguments of M and G^T in the upper lines of (2.10) and (2.11) or to solve the system (2.11) iteratively, if this is computationally advantageous. Convergence of this method is proved in the same way as in the index 2 case.

Velocity Stabilization. It can be seen from (2.6) that errors in the velocity constraint $G(q)u = 0$ are more critical for the numerical solution than errors in the position constraint $g(q) = 0$. It is therefore interesting to study the method, where the numerical solution is projected only to the velocity constraint. Alishenas & Ólafsson (1994) come to the conclusion that “*velocity projection* is the most efficient projection with regard to improvement of the numerical integration”.

We have applied the code DOPRI5 in four different variants to the index 1 formulation of the pendulum equation (2.1): (i) standard application without any projection, (ii) only projection on the position constraint, (iii) only projection on the velocity constraint, (iv) sequential position and velocity projections. The the global

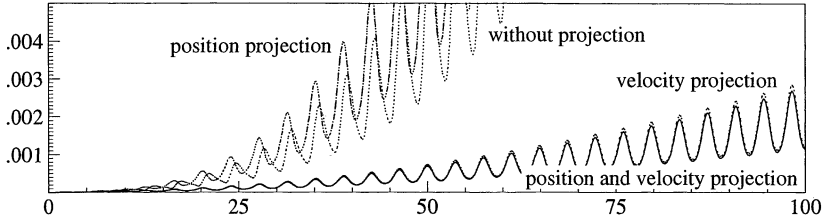


Fig. 2.2. Global error of DOPRI5 with various projections ($Atol = Rtol = 10^{-6}$)

error (in position and velocity) during integration is shown in Fig. 2.2. We conclude that a projection on the position constraint without projection on the velocity constraint does not improve the global error (it makes it even worse in our example). On the other hand, velocity stabilization is as efficient as the complete projection (position and velocity). Nearly no difference can be observed in Fig. 2.2.

Differential Equations with Invariants

Closely related to the above techniques is the numerical treatment of differential equations with invariants. Consider the initial value problem

$$y' = f(y), \quad y(0) = y_0, \quad (2.12)$$

and suppose that the solution is known to have the invariant

$$\varphi(y) = 0. \quad (2.13)$$

For example, the differential equation (1.46a,e) for (q, u) has the invariants (1.46c) and (1.46d). Conservation laws (total energy, ...) may also be written in the form (2.13). The invariant (2.13) is called a *first integral*, if $\varphi_y(y)f(y) \equiv 0$ in a neighbourhood of the solution.

Linear first integrals of the form $\varphi(y) = c + d^T y$ are preserved exactly by most integration methods (e.g., Runge-Kutta and multistep methods). Quadratic first integrals are preserved exactly by symplectic Runge-Kutta methods (see Theorem II.16.7). More complicated invariants are in general not preserved.

The above projection techniques can be adapted to the treatment of the problem (2.12-13) (see Shampine (1986), Eich (1993), Ascher, Chin & Reich (1994)). We apply a numerical method to (2.12) and project (orthogonally or somehow else) the numerical solution onto the manifold defined by (2.13). As discussed above, this procedure retains the order of convergence of the basic method.

Hamiltonian Systems. Differential equations of the form

$$p'_i = -\frac{\partial H}{\partial q_i}(p, q), \quad q'_i = \frac{\partial H}{\partial p_i}(p, q), \quad i = 1, \dots, n, \quad (2.14)$$

where $H : \mathbb{R}^{2n} \rightarrow \mathbb{R}$ is a smooth function, always have $H(p, q) = \text{Const}$ as first integral. It is tempting to exploit this information and project the numerical solution

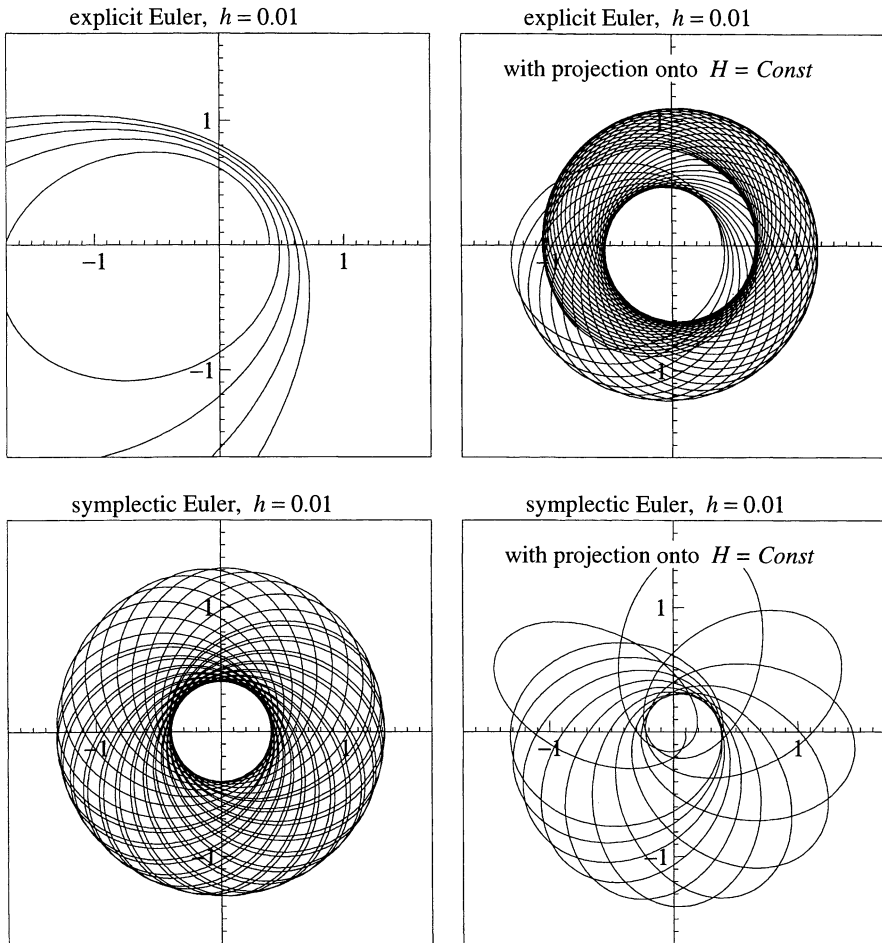


Fig. 2.3. Study of the projection onto the manifold $H(p, q) = H(p_0, q_0)$

onto the manifold $H(p, q) = H(p_0, q_0)$. Consider for example the perturbed Kepler problem with Hamiltonian

$$H(p, q) = \frac{p_1^2 + p_2^2}{2} - \frac{1}{\sqrt{q_1^2 + q_2^2}} - \frac{0.005}{\sqrt{(q_1^2 + q_2^2)^3}} \quad (2.15)$$

and initial values $q_1(0) = 1 - e$, $q_2(0) = 0$, $p_1(0) = 0$, $p_2(0) = \sqrt{(1+e)/(1-e)}$ (eccentricity $e = 0.6$). The upper pictures of Fig. 2.3 show the numerical solution obtained by the explicit Euler method with step size $h = 0.01$; to the left without any projection, and to the right with projection onto $H = \text{Const}$. An improvement can be observed, but the numerical solution still does not reflect the geometric structure of the exact solution (invariant torus). We also have applied the symplectic Euler method (see Eq. (16.54) of Sect. II.16). Here we see that the numerical

solution (without projection) shows the correct qualitative behaviour (this can be explained by a backward error analysis, see Sect. II.16), whereas the projection onto $H = \text{Const}$ destroys this property. A remedy could be the following: apply a symplectic method to the problem, project the numerical solution to $H = \text{Const}$, but continue the integration with the unprojected values.

Methods Based on Local State Space Forms

This method is also called *differential-geometric approach* by Potra & Rheinboldt (1990). The idea is to regard the differential-algebraic system as a differential equation on a manifold (see Sect. VII.1) and to solve the equation in this manifold by introducing suitable local coordinates.

Let us illustrate this approach at the pendulum example. The equations, formulated in cartesian coordinates, are given in the beginning of this section. The solution manifold is (compare with Eq. (1.22))

$$\mathcal{M} = \{(x, y, u, v) \mid x^2 + y^2 = 1, xu + yv = 0\}.$$

This is a 2-dimensional manifold in \mathbb{R}^4 and can be parametrized by (φ, η) as follows:

$$\begin{aligned} x &= \cos \varphi, & u &= -\eta \sin \varphi, \\ y &= \sin \varphi, & v &= \eta \cos \varphi. \end{aligned} \quad (2.16)$$

A short calculation shows that the system (2.1a,b), (2.3), written in the new coordinates, leads to the well-known equation

$$\varphi' = \eta, \quad \eta' = -\cos \varphi. \quad (2.17)$$

This differential equation can be solved numerically without any difficulties. The numerical approximation in the original coordinates is then obtained via (2.16). Obviously, the position and velocity constraints are satisfied exactly.

Although this example nicely illustrates the main ideas, it may be misleading. First of all, in typical applications it is not possible to use one and the same parametrization throughout the whole integration. Secondly, the choice of coordinates is usually not obvious and the transformed differential equation can be much more complicated than the original one (see for example Alishenas (1992)).

Local State Space Form. Suppose that the differential-algebraic system, which we want to solve, can be written as a differential equation

$$y' = v(y), \quad y \in \mathcal{M} \quad (2.18)$$

on a smooth d -dimensional manifold $\mathcal{M} \subset \mathbb{R}^n$. Consider a coordinate function $\omega : U \rightarrow V$ (sufficiently differentiable, bijective, and $\omega'(\eta)$ of full rank) between the open set $U \subset \mathbb{R}^d$ and $V \subset \mathcal{M}$, and denote the coordinates in U by $\eta \in \mathbb{R}^d$. Under the transformation $y = \omega(\eta)$ the equation (2.18) becomes

$$\omega'(\eta)\eta' = v(\omega(\eta)). \quad (2.19)$$

Since $v(y) \in T_y \mathcal{M}$ for all $y \in \mathcal{M}$ (see Eq. (1.19)), there exists η' such that (2.19) holds. Moreover η' is unique, because $\omega'(\eta)$ is of full rank. Using the notation $\omega'(\eta)^+ = (\omega'(\eta)^T \omega'(\eta))^{-1} \omega'(\eta)^T$ for the pseudo-inverse of $\omega'(\eta)$ we therefore obtain

$$\eta' = \omega'(\eta)^+ v(\omega(\eta)), \quad (2.20)$$

which is an ordinary differential equation in \mathbb{R}^d and is called *local state space form* of (2.18). Observe that different coordinate functions lead to different state space forms.

The *numerical procedure* for solving (2.18) is the following: suppose that an approximation $y_k \in \mathcal{M}$ of $y(t_k)$ is given. We then choose a coordinate function and apply a standard method (e.g., Runge-Kutta) with initial value $\eta_k = \omega^{-1}(y_k)$ to the state space form (2.20). This yields an approximation η_{k+1} at time t_{k+1} . Finally, we put $y_{k+1} = \omega(\eta_{k+1})$. By definition of this procedure, the numerical approximation y_{k+1} again lies in \mathcal{M} .

If one uses one and the same local state space form for the whole integration (as it is the case for the pendulum example, Eq. (2.17)), the convergence properties for (2.20) carry immediately over to (2.18) via the coordinate function $y = \omega(\eta)$. In more complex situations it may be necessary to change the coordinates several times, and from a computational point of view it may even be more advantageous to change them in every integration step.

Theorem 2.2. *Consider the above procedure for the numerical solution of (2.18), and denote by $y = \omega_k(\eta)$ the coordinate transformation of the k th step. If, in a neighbourhood of $\omega_k^{-1}(y_k)$, the matrices $\omega'_k(\eta)$ and $\omega'_k(\eta)^+$ are uniformly bounded in k , then the convergence properties for standard ordinary differential equations carry over to the problem (2.18) on a manifold \mathcal{M} .*

Proof. In the case of one-step methods we have

$$y_{k+1} = \omega_k \left(\omega_k^{-1}(y_k) + h \Phi_k(\omega_k^{-1}(y_k), h) \right),$$

where $\Phi_k(\eta, h)$ is the increment function of the method when applied to (2.20) with ω replaced by ω_k . Due to the regularity assumptions on $\omega_k(\eta)$, this formula can be written as

$$y_{k+1} = y_k + h \Psi_k(y_k, h)$$

and takes the form of a standard one-step method. The assumptions guarantee that the functions Ψ_k have a uniform Lipschitz constant with respect to the first argument. Therefore the convergence proofs of Sect. II.3 apply. For multistep methods the situation is analogous. \square

Choice of Local Coordinates. Let us explain two choices for the constrained mechanical system (2.4), whose solution manifold is given by

$$\mathcal{M} = \{(q, u) \mid g(q) = 0, \quad G(q)u = 0\}. \quad (2.21)$$

Here $q, u \in \mathbb{R}^n$ are generalized coordinates, $g(q) \in \mathbb{R}^m$ and $G(q) = g_q(q)$. The adaptation to other differential-algebraic systems with known solution manifold is more or less straightforward.

Generalized Coordinate Partitioning (Wehage & Haug 1982). Assuming that the Jacobian $G(q)$ has full row rank, there exists a partitioning $q = (\eta, \hat{\eta})$ such that $g_{\hat{\eta}}(\eta, \hat{\eta})$ is invertible ($\eta \in \mathbb{R}^{n-m}$, $\hat{\eta} \in \mathbb{R}^m$). By the Implicit Function Theorem the constraint $g(q) = 0$ can be solved for $\hat{\eta}$ in a neighbourhood of a consistent value $q_0 = (\eta_0, \hat{\eta}_0)$. Hence, there exists a function $\hat{\eta} = h(\eta)$ (defined for η close to η_0) such that $g(\eta, h(\eta)) = 0$. With a corresponding partitioning $u = (\nu, \hat{\nu})$ the velocity constraint becomes $g_{\eta}(\eta, \hat{\eta})\nu + g_{\hat{\eta}}(\eta, \hat{\eta})\hat{\nu} = 0$ and allows us to express $\hat{\nu}$ in terms of η, ν as $\hat{\nu} = k(\eta, \nu)$. A coordinate function is thus given by $\omega(\eta, \nu) = ((\eta, h(\eta)), (\nu, k(\eta, \nu)))$, and the differential equation in these local coordinates is

$$\eta' = \nu, \quad \nu' = \nu'(\omega(\eta, \nu)), \quad (2.22)$$

where $\nu'(q, u)$ collects the ν -components of the solution $u'(q, u)$ of the linear system (1.38e). We emphasize that for a numerical implementation the differential equation (2.22) need not be known analytically. However, a nonlinear system has to be solved each time when the right-hand side of (2.22) has to be evaluated.

Tangent Space Parametrization (Potra & Rheinboldt 1991, Yen 1993). Instead of partitioning the components of q and u we split the vectors $q - q_0$ and $u - u_0$ according to

$$q - q_0 = Q_0\eta + Q_1\hat{\eta}, \quad u - u_0 = Q_0\nu + Q_1\hat{\nu}, \quad (2.23)$$

where the columns of Q_0 form a basis of the tangent space $\{v \mid G(q_0)v = 0\}$ to the manifold $g(q) = 0$, which is completed by the columns of Q_1 to a basis of the whole space. The condition $g(q) = 0$ together with the first relation of (2.23) define (locally) q and $\hat{\eta}$ as functions of η . Similarly, $G(q)u = 0$ and the second relation of (2.23) define u and $\hat{\nu}$ as functions of ν and q . Denoting these relationships by $\hat{\eta} = h(\eta)$, $\hat{\nu} = k(\eta, \nu)$, we get formally the same coordinate function as in the previous example, and the state space form is given by

$$\eta' = \nu, \quad \nu' = Q_0^+ u'(\omega(\eta, \nu)), \quad (2.24)$$

where $Q_0^+ = (Q_0^T Q_0)^{-1} Q_0^T$ is the pseudo-inverse of Q_0 , and $u'(q, u)$ denotes the solution of the linear system (2.5).

The evaluation of $h(\eta)$ requires the solution of a nonlinear system, whose Jacobian is

$$\begin{pmatrix} I & -Q_1 \\ G(q_0) & 0 \end{pmatrix}.$$

This suggests to take $-Q_1 = G^T(q_0)$ or better $-Q_1 = M^{-1}(q_0)G^T(q_0)$, so that simplified Newton iterations lead to linear systems with a matrix that already appears in (2.5). The linear system for the computation of $k(\eta, \nu)$ has the same structure.

Due to the fact that the evaluation of the right-hand side of (2.24) requires the solution of a nonlinear system, the authors of this approach prefer the use of multistep methods which, in general, use less function evaluations than one-step methods. In connection with Runge-Kutta methods, Potra (1995) suggests the use of certain predicted values instead of the exact solutions of these nonlinear systems, and requires that only the approximation at the end of every step lies on the manifold \mathcal{M} . The resulting algorithm is then equivalent to solving the index 1 problem combined with projections onto \mathcal{M} at the end of each step.

Overdetermined Differential-Algebraic Equations

In contrast to the approach at the beginning of this section, where the constraint is replaced by one of its derivatives, we consider the original system and one or more derivatives of the constraints as a unity. For example, the equations of motion of a constrained mechanical system become

$$q' = u \quad (2.25a)$$

$$M(q)u' = f(q, u) - G^T(q)\lambda \quad (2.25b)$$

$$0 = g(q) \quad (2.25c)$$

$$0 = G(q)u \quad (2.25d)$$

$$0 = g_{qq}(q)(u, u) + G(q)M(q)^{-1}(f(q, u) - G^T(q)\lambda). \quad (2.25e)$$

This system is overdetermined, because we are concerned with more equations than unknowns. Nevertheless, it possesses a unique solution, if (1.47) is satisfied and consistent initial values are prescribed.

We illustrate the numerical solution of (2.25) with the BDF method. A formal application (see Sect. VI.2) gives

$$q_k - \hat{q} - h\gamma u_k = 0 \quad (2.26a)$$

$$M(q_k)(u_k - \hat{u}) - h\gamma(f(q_k, u_k) - G^T(q_k)\lambda_k) = 0 \quad (2.26b)$$

$$g(q_k) = 0 \quad (2.26c)$$

$$G(q_k)u_k = 0 \quad (2.26d)$$

$$g_{qq}(q_k)(u_k, u_k) + G(q_k)M(q_k)^{-1}(f(q_k, u_k) - G^T(q_k)\lambda_k) = 0, \quad (2.26e)$$

where $\gamma = \beta_k/\alpha_k$, $\hat{q} = (\sum_{i=0}^{k-1} \alpha_i q_i)/\alpha_k$, and $\hat{u} = (\sum_{i=0}^{k-1} \alpha_i u_i)/\alpha_k$ are known quantities. The system (2.26) is overdetermined and does not have a solution, in general. A natural idea (Führer 1988) is to search for a least square solution of (2.26). There are several ways to do this. One can consider different norms, or one can require some of the equations to be exactly satisfied and the remaining ones in a least square sense. Führer & Leimkuhler (1991) impose all constraints (2.26c,d,e), and treat the remaining equations by the use of a special pseudoinverse. This can be achieved by introducing Lagrange multipliers μ_k, η_k in the first two equations

of (2.26) as follows:

$$M(q_k)(q_k - \hat{q} - h\gamma u_k) + h\gamma(G^T(q_k)\mu_k + (G_q(q_k)u_k)^T\eta_k) = 0 \quad (2.27a)$$

$$M(q_k)(u_k - \hat{u}) - h\gamma(f(q_k, u_k) - G^T(q_k)\lambda_k) + h\gamma G^T(q_k)\eta_k = 0. \quad (2.27b)$$

For sufficiently small h , the system (2.27a,b), (2.26c,d,e) has a locally unique solution, if (1.47) is satisfied.

Connection with GGL-Formulation. If we omit the acceleration constraint (2.26e), there is no need for two Lagrange multipliers, and we can put $\eta_k = 0$. The resulting system (2.27a,b), (2.26c,d) is then nothing else than the standard BDF discretization of the system (1.48).

Unstructured Higher Index Problems

We consider a general differential-algebraic system

$$F(u', u) = 0. \quad (2.28)$$

For its numerical solution we shall construct an ‘underlying ODE’ (see Definition 1.2) and solve it by any integration method. This approach has been developed in several papers by Campbell (1989, 1993). We shall explain the main ideas following the presentation of Campbell & Moore (1995).

Inspired by the definition of the differentiation index we consider the *derivative array equations*

$$F(u', u) = 0, \quad \frac{dF(u', u)}{dx} = 0, \quad \dots, \quad \frac{d^m F(u', u)}{dx^m} = 0$$

which we write in compact form as

$$G(u', w, u) = 0, \quad (2.29)$$

where $w = (u'', u''', \dots, u^{(m+1)})$ collects the higher derivatives of u . In Eq. (2.29) we consider w, u , and also u' as independent variables. Besides the usual differentiability assumptions we assume that

- (A1) the matrix $(G_{u'}, G_w)$ is 1-full with respect to u' ; this means that the relation $G_{u'}\Delta u' + G_w\Delta w = 0$ implies $\Delta u' = 0$;
- (A2) the matrix $(G_{u'}, G_w)$ has constant rank;
- (A3) the matrix $(G_{u'}, G_w, G_u)$ has full row rank.

These assumptions are required to hold in a neighbourhood of a particular solution of (2.28). The construction of the underlying ODE is based on the following lemma and on its proof.

Lemma 2.3 (Campbell & Moore 1995). *Consider a sufficiently smooth problem (2.28) and assume that (A1), (A2), and (A3) hold. Then there exist coordinate partitions $w = (w_a, w_b)$, $u = (u_a, u_b)$ (and also $u' = (u'_a, u'_b)$ with the same partition*

as for u), such that the derivative array equations (2.29) are equivalent to

$$\begin{aligned} u'_a &= f_a(u_b), & w_a &= \varphi_2(w_b, u_b) \\ u'_b &= f_b(u_b), & u_a &= \varphi_3(u_b) \end{aligned} \quad (2.30)$$

in a neighbourhood of the consistent initial value (u'_0, w_0, u_0) .

Proof. We consider the matrix $(G_{u'}, G_w, G_u)$ evaluated at (u'_0, w_0, u_0) and perform a QR factorization, where column permutations are restricted to components within the vectors u' , w , and u . This yields

$$Q^T(G_{u'}, G_w, G_u)P = \left(\begin{array}{c|cc|cc} B_1 & C_1 & C_2 & D_1 & D_2 \\ 0 & C_3 & C_4 & D_3 & D_4 \\ 0 & 0 & 0 & D_5 & D_6 \end{array} \right), \quad (2.31)$$

where B_1, C_3, D_5 are nonsingular by Assumption (A3), Q is an orthogonal matrix, and $P = \text{diag}(P_1, P_2, P_3)$ with suitable permutation matrices P_1, P_2, P_3 . Fixing the permutation P , we apply the above factorization also to $(G_{u'}, G_w, G_u)$ evaluated at an arbitrary point (u', w, u) close to (u'_0, w_0, u_0) . Because of Assumption (A2) this gives (2.31) with smooth matrices Q, B_i, C_i , and D_i . The decomposition (2.31) defines the partitions $w = (w_a, w_b)$ and $u = (u_a, u_b)$. The first, second and fourth block-columns in (2.31) form an invertible matrix. The Implicit Function Theorem thus implies that (2.29) can be solved for u', w_a, u_a , and we obtain the equivalent system

$$u' = \varphi_1(w_b, u_b), \quad w_a = \varphi_2(w_b, u_b), \quad u_a = \varphi_3(w_b, u_b).$$

We still have to show that the functions φ_1 and φ_3 are independent of w_b . By definition of the φ_i we have

$$G\left(\varphi_1(w_b, u_b), (\varphi_2(w_b, u_b), w_b), (\varphi_3(w_b, u_b), u_b)\right) = 0.$$

Differentiating with respect to w_b yields

$$G_{u'} \cdot \frac{\partial \varphi_1}{\partial w_b} + G_{w_a} \cdot \frac{\partial \varphi_2}{\partial w_b} + G_{w_b} + G_{u_a} \cdot \frac{\partial \varphi_3}{\partial w_b} = 0. \quad (2.32)$$

Multiplying this relation by Q^T , we see from Eq. (2.31) that $D_5(\partial \varphi_3 / \partial w_b) = 0$. Since D_5 is nonsingular, this implies $(\partial \varphi_3 / \partial w_b) = 0$, so that φ_3 is independent of w_b . Assumption (A1) now implies from (2.32) that also $(\partial \varphi_1 / \partial w_b)$ vanishes. This completes the proof of the lemma. \square

Suppose that we know how to compute $f_a(u_b)$, $f_b(u_b)$ and $\varphi_3(u_b)$ for a given value u_b . From (2.30) we then have an ordinary differential equation for u_b , which can be solved by any integration method (Runge-Kutta or multistep, explicit or implicit, ...), and the remaining components are given by $u_a = \varphi_3(u_b)$. The numerical solution of this method thus preserves all constraints (also the hidden ones).

Computation of the Values $f_a(u_b)$, $f_b(u_b)$ and $\varphi_3(u_b)$. It follows from Assumption (A3) that $(G_{u'}, G_w, G_u)^T G = 0$ is equivalent to $G = 0$. Thus, for given u_b , any method of finding the minimum (u', w, u_a) of the function $G^T G$ may be used. Campbell & Moore (1995) propose the use of Gauss-Newton iterations.

Remark. A closely related algorithm has been proposed by Kunkel & Mehrmann (1996). Instead of extracting from the derivative array equations an ordinary differential equation for all variables, they extract an equivalent index 1 problem and solve it by standard integration methods. This modification usually requires one differentiation less of the original system (2.28).

Exercises

1. Repeat the experiment of Fig. 2.1 with other numerical methods (explicit Euler method, multistep methods, constant and variable step sizes, ...). You will observe that in some situations the error in $g(q_n)$ grows only linearly, and the error in $G(q_n)u_n$ remains bounded. Try to explain this observation.
2. a) Prove that the matrix in (2.5) is 1-full with respect to u' if and only if the restriction of M to the kernel of G is injective (this is exactly the condition that is needed in order to be able to apply the methods of this section).
 b) Show by examples that neither M needs to be nonsingular nor G has to be of full rank in order that the condition of part (a) is satisfied.

VII.3 Multistep Methods for Index 2 DAE

BDF is so beautiful that it is hard to imagine something else could be better.
(L. Petzold 1988, heard by P. Deuffhard)

Convergence results of multistep methods for problems of index at least 2 are harder to obtain than for semi-explicit index 1 problems (see Section VI.2). A first convergence result for BDF schemes, valid for linear constant coefficient DAE's of arbitrary index, was given by Sincovec, Erisman, Yip & Epton (1981). Convergence of BDF for nonlinear DAE systems was then studied by Gear, Gupta & Leimkuhler (1985), Lötstedt & Petzold (1986) and Brenan & Engquist (1988). An independent convergence analysis was given by Griepentrog & März (1986), März (1990). They considered general linear multistep methods and problems, where the differential and algebraic equations (and/or variables) are not explicitly separated.

There are several implementations of the BDF schemes for differential-algebraic systems. The most widely used code is DASSL of Petzold (1982). It is described in detail in the book of Brenan, Campbell & Petzold (1989). Further implementations are LSODI of Hindmarsh (1980) and SPRINT of Berzins & Furzeland (1985).

In this section we consider semi-explicit problems

$$\begin{aligned} y' &= f(y, z) \\ 0 &= g(y). \end{aligned} \tag{3.1}$$

We assume that f and g are sufficiently differentiable and that

$$g_y(y)f_z(y, z) \quad \text{is invertible} \tag{3.2}$$

in a neighbourhood of the solution, so that the problem has index 2. A linear multistep method for (3.1) reads

$$\sum_{i=0}^k \alpha_i y_{n+i} = h \sum_{i=0}^k \beta_i f(y_{n+i}, z_{n+i}) \tag{3.3a}$$

$$0 = g(y_{n+k}). \tag{3.3b}$$

This is not the only meaningful definition of a multistep method for (3.1). One could as well replace (3.3b) by

$$0 = \sum_{i=0}^k \beta_i g(y_{n+i}), \tag{3.4}$$

which is obtained by putting $\varepsilon = 0$ in (VI.2.1). The following results can be extended without any difficulty to the second approach. For BDF schemes (where $\beta_0 = \dots = \beta_{k-1} = 0$) both definitions are equivalent.

The convergence results of this section are also valid for index 2 systems of the form $y' = f(y, z)$, $0 = g(y, z)$, if they can be transformed to (3.1) without any differentiation (see the discussion after Eq. (1.14)). This is because the multistep method (3.3) is invariant with respect to these transformations. The same is true for problems of the form $M(u)u' = \varphi(u)$, if the multistep method is defined by

$$\sum_{i=0}^k \alpha_i u_{n+i} = h \sum_{i=0}^k \beta_i v_{n+i}, \quad M(u_{n+k})v_{n+k} = \varphi(u_{n+k}). \quad (3.5)$$

Existence and Uniqueness of Numerical Solution

Equations (3.3) constitute a nonlinear system for y_{n+k}, z_{n+k} . We have the following result about the existence of its solution.

Theorem 3.1. *Suppose that for a solution $y(x), z(x)$ of (3.1) the starting values satisfy for $j = 0, \dots, k-1$ and $x_j = x_0 + jh$*

$$y_j - y(x_j) = \mathcal{O}(h), \quad z_j - z(x_j) = \mathcal{O}(h), \quad g(y_j) = \mathcal{O}(h^2). \quad (3.6)$$

If (3.2) holds in a neighbourhood of this solution and if $\beta_k \neq 0$, then the nonlinear system

$$\sum_{i=0}^k \alpha_i y_i = h \sum_{i=0}^k \beta_i f(y_i, z_i) \quad (3.7a)$$

$$0 = g(y_k) \quad (3.7b)$$

has a solution for $h \leq h_0$. This solution is locally unique and satisfies

$$y_k - y(x_k) = \mathcal{O}(h), \quad z_k - z(x_k) = \mathcal{O}(h). \quad (3.8)$$

Proof. We put

$$\eta = - \sum_{i=0}^{k-1} \frac{\alpha_i}{\alpha_k} y_i + h \sum_{i=0}^{k-1} \frac{\beta_i}{\alpha_k} f(y_i, z_i) \quad (3.9)$$

and define ζ close to $z(x_k)$ such that $g_y(\eta)f(\eta, \zeta) = 0$. We further replace $h(\beta_k/\alpha_k)$ by a new step size which we again denote by h . Then the system (3.7) is equivalent to

$$y_k = \eta + hf(y_k, z_k) \quad (3.10a)$$

$$0 = g(y_k) \quad (3.10b)$$

which is simply the implicit Euler method.

We next show that

$$\eta - y(x_k) = \mathcal{O}(h), \quad \zeta - z(x_k) = \mathcal{O}(h), \quad g(\eta) = \mathcal{O}(h^2). \quad (3.11)$$

The first relation follows from $y_j - y(x_j) = \mathcal{O}(h)$ and from $\sum_{i=0}^k \alpha_i = 0$; the second is a consequence of the definition of ζ and of (3.2). The last relation of (3.11) can be seen as follows: we replace all $f(y_i, z_i)$ in (3.9) by $f(y(x_k), z(x_k))$, introducing an error of size $\mathcal{O}(h^2)$ in η . Hence

$$\eta - y(x_k) = - \sum_{i=0}^{k-1} \frac{\alpha_i}{\alpha_k} (y_i - y(x_k)) + h \left(\sum_{i=0}^{k-1} \frac{\beta_i}{\alpha_k} \right) f(y(x_k), z(x_k)) + \mathcal{O}(h^2).$$

Because of (1.14b,c) this implies

$$g(\eta) = - \sum_{i=0}^{k-1} \frac{\alpha_i}{\alpha_k} g_y(y(x_k)) (y_i - y(x_k)) + \mathcal{O}(h^2). \quad (3.12)$$

The last statement of (3.11) now follows from the fact that $g_y(y(x_k))(y_i - y(x_k)) = g(y_i) + \mathcal{O}(h^2)$ and from (3.6).

To show the existence of a locally unique solution of (3.10), it is possible to adapt the proof of “Theorem 4.1” of HLR89 to the implicit Euler method. We shall, however, reformulate (3.10) in such a way that the implicit function theorem is applicable. We write (3.10b) as

$$\begin{aligned} 0 &= g(y_k) = g(y_k) - g(\eta(h)) + g(\eta(h)) \\ &= \int_0^1 g_y \left(\eta(h) + \tau(y_k - \eta(h)) \right) d\tau \cdot (y_k - \eta(h)) + g(\eta(h)) \end{aligned} \quad (3.13)$$

where we have explicitly indicated the dependence of η on h . Replacing the factor $y_k - \eta(h)$ by $hf(y_k, z_k)$ from (3.10a) and dividing by h we get the system

$$y_k - \eta(h) - hf(y_k, z_k) = 0 \quad (3.14a)$$

$$\int_0^1 g_y \left(\eta(h) + \tau(y_k - \eta(h)) \right) d\tau \cdot f(y_k, z_k) + \frac{1}{h} g(\eta(h)) = 0 \quad (3.14b)$$

which is the discrete analogue of system (1.14a,c). For $h = 0$ the values $y_k = \eta(0)$ and $z_k = \zeta(0)$ satisfy (3.14) because $g(\eta(h)) = \mathcal{O}(h^2)$ and $g_y(\eta)f(\eta, \zeta) = 0$. Further, the derivative of (3.14) with respect to (y_k, z_k) is of the form

$$\begin{pmatrix} I + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & (g_y f_z)(\eta, \zeta) + \mathcal{O}(h) \end{pmatrix}, \quad (3.15)$$

which has a bounded inverse for $h \leq h_0$. Therefore the implicit function theorem (Ortega & Rheinboldt 1970, p. 128) yields the existence of a locally unique solution of (3.14) and hence also of (3.10) and (3.7). \square

Influence of Perturbations

The influence of perturbations in the multistep formula (3.3) on the numerical solution will be studied in the next theorem.

Theorem 3.2. *Let y_k, z_k be given by (3.7) and consider perturbed values \hat{y}_k, \hat{z}_k satisfying*

$$\sum_{i=0}^k \alpha_i \hat{y}_i = h \sum_{i=0}^k \beta_i f(\hat{y}_i, \hat{z}_i) + h\delta \quad (3.16a)$$

$$0 = g(\hat{y}_k) + \theta. \quad (3.16b)$$

In addition to the assumptions of Theorem 3.1 suppose that for $j = 0, \dots, k-1$

$$\hat{y}_j - y_j = \mathcal{O}(h^2), \quad \hat{z}_j - z_j = \mathcal{O}(h), \quad \delta = \mathcal{O}(h), \quad \theta = \mathcal{O}(h^2). \quad (3.17)$$

Then, for $h \leq h_0$ we have the estimates

$$\begin{aligned} \|\hat{y}_k - y_k\| &\leq C \left(\|\hat{Y}_0 - Y_0\| + h\|\hat{Z}_0 - Z_0\| + h\|\delta\| + \|\theta\| \right) \\ \|\hat{z}_k - z_k\| &\leq \frac{C}{h} \left(\sum_{j=0}^{k-1} \|g_y(\hat{y}_k)(\hat{y}_j - y_j)\| + h\|\hat{Y}_0 - Y_0\| \right. \\ &\quad \left. + h\|\hat{Z}_0 - Z_0\| + h\|\delta\| + \|\theta\| \right) \end{aligned} \quad (3.18)$$

where $\hat{Y}_0 - Y_0 = (\hat{y}_{k-1} - y_{k-1}, \dots, \hat{y}_0 - y_0)^T$, $\|\hat{Y}_0 - Y_0\| = \max_{0 \leq j \leq k-1} \|\hat{y}_j - y_j\|$, and likewise for the z -component.

Proof. In analogy to the proof of Theorem 3.1 we put

$$\hat{\eta} = - \sum_{i=0}^{k-1} \frac{\alpha_i}{\alpha_k} \hat{y}_i + h \sum_{i=0}^{k-1} \frac{\beta_i}{\alpha_k} f(\hat{y}_i, \hat{z}_i)$$

and rescale h and δ , so that (3.16) becomes

$$\hat{y}_k = \hat{\eta} + h f(\hat{y}_k, \hat{z}_k) + h\delta \quad (3.19a)$$

$$0 = g(\hat{y}_k) + \theta. \quad (3.19b)$$

As in the proof of Theorem 3.1 we conclude from (3.17) that $\hat{y}_k - \hat{\eta} = \mathcal{O}(h)$ and $\hat{z}_k - \hat{\zeta} = \mathcal{O}(h)$, where $\hat{\zeta}$ is such that $g_y(\hat{\eta})f(\hat{\eta}, \hat{\zeta}) = 0$. Inspired by Eq. (3.14) we rewrite (3.19b) as

$$0 = \int_0^1 g_y \left(\hat{\eta} + \tau (\hat{y}_k - \hat{\eta}) \right) d\tau \cdot (f(\hat{y}_k, \hat{z}_k) + \delta) + \frac{1}{h} g(\hat{\eta}) + \frac{1}{h} \theta, \quad (3.20)$$

which is now a discrete analogue of Eq. (1.29). Subtracting (3.20) from (3.14b) and

exploiting the fact that the matrix $g_y f_z$ is invertible, we deduce the estimate

$$\|\widehat{z}_k - z_k\| \leq C \left(\|\widehat{y}_k - y_k\| + \|\widehat{\eta} - \eta\| + \|\delta\| + \frac{1}{h} \|g(\widehat{\eta}) - g(\eta)\| + \frac{1}{h} \|\theta\| \right). \quad (3.21)$$

A Lipschitz condition for f applied to the difference of (3.19a) and (3.14a) yields

$$\|\widehat{y}_k - y_k\| \leq \|\widehat{\eta} - \eta\| + hL(\|\widehat{y}_k - y_k\| + \|\widehat{z}_k - z_k\|) + h\|\delta\|.$$

Combining the last two estimates we get

$$\begin{aligned} \|\widehat{y}_k - y_k\| &\leq C(\|\widehat{\eta} - \eta\| + h\|\delta\| + \|\theta\|) \\ \|\widehat{z}_k - z_k\| &\leq \frac{C}{h} (\|g_y(\widehat{\eta})(\widehat{\eta} - \eta)\| + h\|\widehat{\eta} - \eta\| + h\|\delta\| + \|\theta\|). \end{aligned} \quad (3.22)$$

The conclusion now follows from the definitions of η and ζ and from $\widehat{y}_k - \widehat{\eta} = \mathcal{O}(h)$. \square

Remark 3.3. a) The above proof shows that the constant C in (3.18) depends on bounds for certain derivatives of f and g , but not on the constants implied by the $\mathcal{O}(\dots)$ terms in (3.17) (if h is sufficiently small). This observation will be used in the convergence proof below.

b) For one-step methods (e.g., implicit Euler method, trapezoidal rule) the term $\|\sum_{j=0}^{k-1} g_y(\widehat{y}_k)(\widehat{y}_j - y_j)\|$ can be omitted in (3.18), if we require $g(y_0) = g(\widehat{y}_0) = 0$. Indeed, it follows from $\widehat{y}_1 = \widehat{y}_0 + \mathcal{O}(h)$ that $g_y(\widehat{y}_1)(\widehat{y}_0 - y_0) = g_y(\widehat{y}_0)(\widehat{y}_0 - y_0) + \mathcal{O}(h\|\widehat{y}_0 - y_0\|)$. Further we have

$$g_y(\widehat{y}_0)(\widehat{y}_0 - y_0) = g(\widehat{y}_0) - g(y_0) + \mathcal{O}(\|\widehat{y}_0 - y_0\|^2),$$

so that the term in question is estimated by $\mathcal{O}(h\|\widehat{y}_0 - y_0\|)$ if h is sufficiently small.

The Local Error

Consider initial values $y_j = y(x_j)$, $z_j = z(x_j)$ ($j = 0, \dots, k-1$) on the exact solution of (3.1) and apply the multistep formula (3.7) once. The differences $y_k - y(x_k)$ and $z_k - z(x_k)$ are then called the *local errors* of the method.

Lemma 3.4. *Suppose that the DAE (3.1) satisfies (3.2) and that the multistep method (3.7) has order p (in the sense of Sect. III.2). Then its local error satisfies*

$$y_k - y(x_k) = \mathcal{O}(h^{p+1}), \quad z_k - z(x_k) = \mathcal{O}(h^p). \quad (3.23)$$

Proof. We put $\widehat{y}_j = y(x_j)$, $\widehat{z}_j = z(x_j)$ for $j = 0, \dots, k$. These values satisfy (3.16) with $\delta = \mathcal{O}(h^p)$ and $\theta = 0$. Since $\widehat{y}_j = y_j$ and $\widehat{z}_j = z_j$ for $j < k$, the statement follows immediately from Theorem 3.2. \square

Convergence for BDF

The study of convergence is simpler for BDF schemes than for general multi-step methods, because y_{n+k} depends only on y_n, \dots, y_{n+k-1} , but not on z_n, \dots, z_{n+k-1} (due to $\beta_0 = \dots = \beta_{k-1} = 0$). Therefore the y - and z -components can be treated separately. The following convergence result was obtained by Gear, Gupta & Leimkuhler (1985), Lötstedt & Petzold (1986) and Brenan & Engquist (1988).

Theorem 3.5. *Consider an index 2 problem (3.1) which satisfies (3.2). Then the k -step BDF scheme (III.1.22') is convergent of order $p = k$, if $k \leq 6$; i.e.,*

$$y_n - y(x_n) = \mathcal{O}(h^p), \quad z_n - z(x_n) = \mathcal{O}(h^p) \quad \text{for } x_n = nh \leq \text{Const}, \quad (3.24)$$

whenever the initial values satisfy

$$y_j - y(x_j) = \mathcal{O}(h^{p+1}) \quad \text{for } j = 0, \dots, k-1. \quad (3.25)$$

Remark. The assumption (3.25) can be relaxed to $y_j - y(x_j) = \mathcal{O}(h^p)$ for $k \geq 3$, but not for $k = 1$ (see Exercise 1).

Proof. We combine the convergence proof for Runge-Kutta methods (HLR89, Theorem 4.4) with the techniques of Sect. III.4. Inspired by Lady Windermere's Fan (Fig. III.4.1) we first study the propagation of the local errors and their accumulation over the whole interval for the y -component (part a). The z -component is treated in part (b) and technical details are given in part (c).

a) In addition to the numerical solution $\{y_n, z_n\}$, which we now also denote by $\{y_n^0, z_n^0\}$, we consider for $\ell = 1, 2, \dots$ the multistep solutions $\{y_n^\ell, z_n^\ell\}$ with starting values $y_j^\ell = y(x_j)$, $z_j^\ell = z(x_j)$ for $j = \ell-1, \dots, \ell+k-2$ on the exact solution. Our first aim is to estimate $y_n^\ell - y_n^{\ell+1}$ in terms of the local errors $y_{\ell+k-1}^\ell - y_{\ell+k-1}^{\ell+1}$ (or starting errors if $\ell = 0$). For simplicity we omit the upper index and consider two neighbouring multistep solutions $\{\hat{y}_n, \hat{z}_n\}$ and $\{\tilde{y}_n, \tilde{z}_n\}$. In order to be able to apply Theorem 3.2 we fix three sufficiently large constants C_0, C_1, C_2 and suppose that for $nh \leq \text{Const}$

$$\|\hat{y}_n - y(x_n)\| \leq C_0 h, \quad \|\tilde{y}_n - \hat{y}_n\| \leq C_1 h^2, \quad \|\hat{z}_n - z(x_n)\| \leq C_2 h. \quad (3.26)$$

This will be justified in part (c) below. We introduce the notation $\Delta y_n = \tilde{y}_n - \hat{y}_n$, $\Delta z_n = \tilde{z}_n - \hat{z}_n$ and $\Delta Y_n = (\Delta y_{n+k-1}, \dots, \Delta y_n)^T$. Observing that y_{n+k}, z_{n+k} do not depend on z_n, \dots, z_{n+k-1} for the BDF schemes, it follows from Theorem 3.2 with $\delta = 0$ and $\theta = 0$ that

$$\|\Delta y_{n+k}\| \leq C \|\Delta Y_n\| \quad (3.27a)$$

$$\|\Delta z_{n+k}\| \leq \frac{C}{h} \left(\sum_{j=0}^{k-1} \|g_y(\hat{y}_{n+k}) \Delta y_{n+j}\| + h \|\Delta Y_n\| \right). \quad (3.27b)$$

Here C does not depend on the choice of C_0, C_1, C_2 , if h is sufficiently small (see Remark 3.3a). Our assumption (3.26) together with (3.27) implies $\Delta y_{n+k} = \mathcal{O}(h^2)$

and $\Delta z_{n+k} = \mathcal{O}(h)$. We therefore obtain by linearization of the multistep formula

$$\sum_{i=0}^k \alpha_i \Delta y_{n+i} = h \beta_k f_z(\widehat{y}_{n+k}, \widehat{z}_{n+k}) \Delta z_{n+k} + \mathcal{O}(h \|\Delta Y_n\|) \quad (3.28a)$$

$$0 = g_y(\widehat{y}_{n+k}) \Delta y_{n+k} + \mathcal{O}(h \|\Delta Y_n\|). \quad (3.28b)$$

We next use the projections (see also Definition 4.3 below)

$$Q_n = (f_z(g_y f_z)^{-1} g_y)(\widehat{y}_{n+k}, \widehat{z}_{n+k}), \quad P_n = I - Q_n \quad (3.29)$$

for which

$$P_n^2 = P_n, \quad Q_n^2 = Q_n, \quad P_n Q_n = Q_n P_n = 0, \quad Q_{n+1} = Q_n + \mathcal{O}(h). \quad (3.30)$$

The last relation of (3.30) follows from (3.26) and the smoothness of the solution $y(x), z(x)$. We then multiply (3.28a) by P_{n+k} (which eliminates Δz_{n+k}) and (3.28b) by $f_z(g_y f_z)^{-1}$. This yields with (3.30)

$$\sum_{i=0}^k \alpha_i P_{n+i} \Delta y_{n+i} = \mathcal{O}(h \|\Delta Y_n\|), \quad Q_{n+k} \Delta y_{n+k} = \mathcal{O}(h \|\Delta Y_n\|). \quad (3.31)$$

Introducing the vectors

$$U_n = (P_{n+k-1} \Delta y_{n+k-1}, \dots, P_n \Delta y_n)^T, \\ V_n = (Q_{n+k-1} \Delta y_{n+k-1}, \dots, Q_n \Delta y_n)^T,$$

we have $\Delta Y_n = U_n + V_n$ and the relations (3.31) become

$$U_{n+1} = (A \otimes I) U_n + \mathcal{O}(h \|U_n\| + h \|V_n\|) \quad (3.32a)$$

$$V_{n+1} = (N \otimes I) V_n + \mathcal{O}(h \|U_n\| + h \|V_n\|) \quad (3.32b)$$

where (with $\alpha'_j = \alpha_j / \alpha_k$)

$$A = \begin{pmatrix} -\alpha'_{k-1} & \dots & -\alpha'_1 & -\alpha'_0 \\ 1 & & 0 & 0 \\ & \ddots & \vdots & \vdots \\ & & 1 & 0 \end{pmatrix}, \quad N = \begin{pmatrix} 0 & \dots & 0 & 0 \\ 1 & & 0 & 0 \\ & \ddots & \vdots & \vdots \\ & & 1 & 0 \end{pmatrix}. \quad (3.33)$$

According to Lemma III.4.4 we now choose a norm $\|U\|$ such that $\|A \otimes I\| \leq 1$. We then choose a (possibly different) norm $\|V\|$, for which $\|N \otimes I\| \leq \varrho < 1$. Consequently it follows from (3.32) that

$$\begin{pmatrix} \|U_{n+1}\| \\ \|V_{n+1}\| \end{pmatrix} \leq \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(h) & \varrho + \mathcal{O}(h) \end{pmatrix} \begin{pmatrix} \|U_n\| \\ \|V_n\| \end{pmatrix}. \quad (3.34)$$

As in the proof of Lemma VI.3.9 we diagonalize the matrix in (3.34) and so obtain

$$\begin{aligned} \|\Delta Y_n\| &\leq \text{Const}_1 (\|U_n\| + \|V_n\|) \\ &\leq \text{Const}_2 (\|U_0\| + (\varrho^n + h) \|V_0\|), \end{aligned} \quad (3.35a)$$

$$\|V_n\| \leq \text{Const}_3 (h \|U_0\| + (\varrho^n + h) \|V_0\|). \quad (3.35b)$$

The vectors U_0 and V_0 are composed of local errors (of the y -component) or of errors in the starting values, which are of size $\mathcal{O}(h^{p+1})$ by (3.23) and (3.25). Hence, it follows from (3.35) that the propagated errors satisfy

$$\begin{aligned} \|\Delta y_n\| &\leq C_3 h^{p+1}, \\ \|g_y(\hat{y}_{n+k})\Delta y_{n+j}\| &\leq C_4(\varrho^n + h)h^{p+1} \quad \text{for } j = 0, \dots, k-1. \end{aligned} \quad (3.36)$$

Summing up we obtain

$$\|y_n - y(x_n)\| \leq \sum_{\ell=0}^{n-k+1} \|y_n^\ell - y_n^{\ell+1}\| \leq C_5 h^p, \quad (3.37)$$

the desired estimate for the y -component.

b) Since z_n depends only on y_{n-k}, \dots, y_{n-1} but not on the previous z -values, we can apply Theorem 3.2 with $\hat{y}_i = y(x_i)$, $\hat{z}_i = z(x_i)$, $\delta = \mathcal{O}(h^p)$ and $\theta = 0$. This yields

$$\|z_n - z(x_n)\| \leq \frac{C}{h} \sum_{j=1}^k \|g_y(y(x_n))(y_{n-j} - y(x_{n-j}))\| + \mathcal{O}(h^p). \quad (3.38)$$

Using (3.36) and $y_n^\ell = y(x_n) + \mathcal{O}(h^p)$, which follows as in (3.37), we obtain

$$\begin{aligned} \|g_y(y(x_n))(y_{n-j} - y(x_{n-j}))\| &= \left\| \sum_{\ell=0}^{n-k+1} g_y(y(x_n))(y_{n-j}^\ell - y_{n-j}^{\ell+1}) \right\| \\ &\leq \sum_{\ell=0}^{n-k+1} \left(\|g_y(y_n^\ell)(y_{n-j}^\ell - y_{n-j}^{\ell+1})\| + \mathcal{O}(h^{2p+1}) \right) = \mathcal{O}(h^{p+1}) \end{aligned}$$

and hence also

$$\|z_n - z(x_n)\| \leq C_6 h^p. \quad (3.39)$$

c) In general, the constants C_3 , C_5 and C_6 will depend on C_0, C_1, C_2 of our assumption (3.26). For $p \geq 2$ we can restrict the step size h so that

$$C_5 h^{p-1} \leq C_0, \quad C_3 h^{p-1} \leq C_1, \quad C_6 h^{p-1} \leq C_2$$

and the numerical solutions will never violate the conditions (3.26) on the considered interval.

For $p = 1$ (the implicit Euler method) we know from Remark 3.3b that the estimate (3.27b) can be replaced by

$$\|\Delta z_{n+k}\| \leq C \|\Delta Y_n\|. \quad (3.40)$$

Instead of (3.28a) we thus immediately get

$$\Delta y_{n+1} - \Delta y_n = \mathcal{O}(h \|\Delta y_n\|) \quad (3.41)$$

where the constant implied by the $\mathcal{O}(\dots)$ term is independent of C_0, C_1, C_2 , if h is sufficiently small. Standard techniques (without considering the projections (3.29)) then yield the convergence result. \square

With the ideas of Sect. III.5 the above proof can be extended to cover variable step sizes as well. Originally, such a convergence result was given by Gear, Gupta & Leimkuhler (1985).

General Multistep Methods

For a general multistep method (3.3) with generating polynomials

$$\varrho(\zeta) = \sum_{i=0}^k \alpha_i \zeta^i, \quad \sigma(\zeta) = \sum_{i=0}^k \beta_i \zeta^i$$

we have the following convergence result.

Theorem 3.6. *Consider an index 2 problem (3.1) which satisfies (3.2). Assume that the multistep method is stable (Definition III.3.2) and strictly stable at infinity (the zeros of $\sigma(\zeta)$ lie inside the unit disc $|\zeta| < 1$). If its order is $p \geq 2$, then the global error satisfies*

$$y_n - y(x_n) = \mathcal{O}(h^p), \quad z_n - z(x_n) = \mathcal{O}(h^p) \quad \text{for } x_n = nh \leq \text{Const}$$

whenever the initial values satisfy (for $j = 0, \dots, k-1$)

$$y_j - y(x_j) = \mathcal{O}(h^{p+1}), \quad z_j - z(x_j) = \mathcal{O}(h^p). \quad (3.42)$$

Proof. The proof is essentially the same as for the BDF schemes. Due to the dependence of y_{n+k}, z_{n+k} on y_n, \dots, y_{n+k-1} and on z_n, \dots, z_{n+k-1} the following modifications are necessary.

In addition to (3.26) we assume $\|\tilde{z}_n - \hat{z}_n\| \leq C_3 h$. Instead of (3.27) we have (from Theorem 3.2)

$$\begin{aligned} \|\Delta y_{n+k}\| &\leq C(\|\Delta Y_n\| + h\|\Delta Z_n\|) \\ \|\Delta z_{n+k}\| &\leq \frac{C}{h} \left(\sum_{j=0}^{k-1} \|g_y(\hat{y}_{n+k}) \Delta y_{n+j}\| + h\|\Delta Y_n\| + h\|\Delta Z_n\| \right) \end{aligned}$$

and (3.28) becomes

$$\begin{aligned} \sum_{i=0}^k \alpha_i \Delta y_{n+i} &= h \sum_{i=0}^k \beta_i f_z(\hat{y}_{n+k}, \hat{z}_{n+k}) \Delta z_{n+i} + \mathcal{O}(h\|\Delta Y_n\| + h^2\|\Delta Z_n\|) \\ 0 &= g_y(\hat{y}_{n+k}) \Delta y_{n+k} + \mathcal{O}(h\|\Delta Y_n\| + h^2\|\Delta Z_n\|). \end{aligned} \quad (3.43)$$

A recursion for Δz_n is obtained as follows: we multiply the upper line of (3.43) by $((g_y f_z)^{-1} g_y)(\hat{y}_{n+k}, \hat{z}_{n+k})$ and so get

$$\begin{aligned} h \sum_{i=0}^k \beta_i \Delta z_{n+i} &= \sum_{i=0}^k \alpha_i ((g_y f_z)^{-1} g_y)(\hat{y}_{n+k}, \hat{z}_{n+k}) \Delta y_{n+i} \\ &\quad + \mathcal{O}(h\|\Delta Y_n\| + h^2\|\Delta Z_n\|). \end{aligned} \quad (3.44)$$

With the projections P_n, Q_n of (3.29) and the vectors U_n, V_n we thus obtain (3.32) with an additional $\mathcal{O}(h^2 \|\Delta Z_n\|)$ term. From (3.44) we get

$$h \Delta Z_{n+1} = (B \otimes I) h \Delta Z_n + \mathcal{O}(h \|U_n\| + \|V_n\| + h^2 \|\Delta Z_n\|),$$

where

$$B = \begin{pmatrix} -\beta'_{k-1} & \cdots & -\beta'_1 & -\beta'_0 \\ 1 & & 0 & 0 \\ & \ddots & \vdots & \vdots \\ & & 1 & 0 \end{pmatrix}$$

with $\beta'_j = \beta_j / \beta_k$. For this equation we use a norm for which $\|B \otimes I\| \leq \kappa < 1$. This is possible, because the method is strictly stable at infinity. Summarizing, we get the inequality

$$\begin{pmatrix} \|U_{n+1}\| \\ \|V_{n+1}\| \\ h \|\Delta Z_{n+1}\| \end{pmatrix} \leq \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(h) & \varrho + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(h) & \mathcal{O}(1) & \kappa + \mathcal{O}(h) \end{pmatrix} \begin{pmatrix} \|U_n\| \\ \|V_n\| \\ h \|\Delta Z_n\| \end{pmatrix} \quad (3.45)$$

which can be solved as before and yields

$$\begin{aligned} \|\Delta y_n\| &\leq C_3 h^{p+1}, & \|\Delta z_n\| &\leq C_7 (\varrho^n + \kappa^n + h) h^p, \\ \|g_y(\hat{y}_{n+k}) \Delta y_{n+j}\| &\leq C_4 (\varrho^n + \kappa^n + h) h^{p+1} & \text{for } j = 0, \dots, k-1. \end{aligned} \quad (3.46)$$

Summing up the propagated errors as in (3.37) we obtain the desired estimates for the y - and z -component. \square

Solution of the Nonlinear System by Simplified Newton

The nonlinear system (3.3) is usually solved by a simplified Newton iteration and it is interesting to study its convergence. As in the proof of Theorem 3.1 we introduce η by (3.9) and rescale h so that the nonlinear system becomes (omitting the indices)

$$\begin{aligned} y - \eta - h f(y, z) &= 0 \\ g(y) &= 0. \end{aligned} \quad (3.47)$$

This is just the implicit Euler method and we can apply the discussion of HLR89, Chapter 7. The Jacobian of the nonlinear system (3.47) is

$$J = \begin{pmatrix} I - h f_y & -h f_z \\ g_y & 0 \end{pmatrix} \quad (3.48)$$

and its inverse has the form

$$J^{-1} = \begin{pmatrix} P + \mathcal{O}(h) & f_z(g_y f_z)^{-1} + \mathcal{O}(h) \\ -h^{-1}(g_y f_z)^{-1} g_y + \mathcal{O}(1) & h^{-1}(g_y f_z)^{-1} + \mathcal{O}(1) \end{pmatrix} \quad (3.49)$$

where $P = I - f_z(g_y f_z)^{-1} g_y$ is the projection of (3.29). We now consider the

simplified Newton method as a fixed point iteration with the function

$$\Phi(y, z) = \begin{pmatrix} y \\ z \end{pmatrix} - J_0^{-1} \begin{pmatrix} y - \eta - hf(y, z) \\ g(y) \end{pmatrix}. \quad (3.50)$$

The subscript 0 in J_0 indicates that the arguments of the derivatives in (3.48) are evaluated at some *fixed* approximation $(\hat{\eta}, \hat{\zeta})$ to the solution of (3.47). We shall use the notation $\{f_y\}_0$ for $f_y(\hat{\eta}, \hat{\zeta})$, etc. Direct calculation of $\Phi'(y, z)$ gives

$$\begin{pmatrix} \{f_z(g_y f_z)^{-1}\}_0(\{g_y\}_0 - g_y) + \mathcal{O}(h) & h\{P\}_0 f_z + \mathcal{O}(h^2) \\ h^{-1}\{(g_y f_z)^{-1}\}_0(\{g_y\}_0 - g_y) + \mathcal{O}(1) & \{(g_y f_z)^{-1} g_y\}_0(\{f_z\}_0 - f_z) + \mathcal{O}(h) \end{pmatrix}.$$

If we assume that $(\hat{\eta}, \hat{\zeta})$ approximates the fixed point of (3.50) with an error of $\mathcal{O}(h)$, then we have at this fixed point

$$\Phi'(y, z) = \begin{pmatrix} \mathcal{O}(h) & \mathcal{O}(h^2) \\ \mathcal{O}(1) & \mathcal{O}(h) \end{pmatrix}. \quad (3.51)$$

With the scaling matrix $D = \text{diag}(I, hI)$ (this corresponds to a multiplication of the z -variables by h) we have $\|D\Phi'(y, z)D^{-1}\| = \mathcal{O}(h)$. In the norm $\|y\| + h\|z\|$ we therefore gain a factor h in each simplified Newton iteration.

Remark. The above analysis remains valid if f_y or parts of it are replaced by zero in J_0 . For mechanical problems such an algorithm was proposed by Gear, Gupta & Leimkuhler (1985).

Exercises

1. Show that the assumption $g(y_j) = \mathcal{O}(h^2)$ for $j = 0, \dots, k-1$ cannot be omitted in Theorem 3.1.

Counterexample. Consider the system $x' = 1, y' = k(z), 0 = y - x$, where $k(z) = (e^{z-1} + 1)/2$. Apply the implicit Euler method with initial values $x_0 = 0, y_0 = h, z_0 = 1$.

2. (Gear, Hsu & Petzold 1981, Gear & Petzold 1984). Consider the problem

$$\begin{pmatrix} 0 & 0 \\ 1 & \eta x \end{pmatrix} \begin{pmatrix} y' \\ z' \end{pmatrix} + \begin{pmatrix} 1 & \eta x \\ 0 & 1 + \eta \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} f(x) \\ g(x) \end{pmatrix}. \quad (3.52)$$

- a) Prove that the system (3.52) has index 2 for all values of η .
- b) The z -component of the exact solution is $z(x) = g(x) - f'(x)$.
- c) The implicit Euler method, applied to (3.52) in an obvious manner, yields the recursion

$$z_{n+1} = \frac{\eta}{1+\eta} z_n + \frac{1}{1+\eta} \left(g(x_{n+1}) - \frac{f(x_{n+1}) - f(x_n)}{h} \right).$$

Hence, the method is convergent for $\eta > -1/2$, but unstable for $\eta < -1/2$. For $\eta = -1$ the numerical solution does not exist.

VII.4 Runge-Kutta Methods for Index 2 DAE

RK methods prove popular at IMA conference on numerical ODEs.
(Byrne & Hindmarsh, SIAM News, March 1990)

This section is devoted to the convergence of implicit Runge-Kutta methods for semi-explicit index 2 systems (3.1) which satisfy (3.2). The ε -embedding method of Sect. VI.1 defines the numerical solution by

$$y_{n+1} = y_n + h \sum_{i=1}^s b_i k_{ni}, \quad z_{n+1} = z_n + h \sum_{i=1}^s b_i \ell_{ni} \quad (4.1a)$$

where

$$k_{ni} = f(Y_{ni}, Z_{ni}), \quad 0 = g(Y_{ni}) \quad (4.1b)$$

and the internal stages are given by

$$Y_{ni} = y_n + h \sum_{j=1}^s a_{ij} k_{nj}, \quad Z_{ni} = z_n + h \sum_{j=1}^s a_{ij} \ell_{nj} \quad (4.1c)$$

(the state space form method (VI.1.12) does not make sense here, because the algebraic conditions do not depend on z).

The first convergence results for this situation are due to Petzold (1986). They are formulated for general problems $F(y', y) = 0$ under the assumption of “uniform index one”. Since the system (3.1) becomes “uniform index one” if we replace z by u' (Gear 1988, see also Exercise 1), the results of Petzold can be applied to (3.1). A further study for the semi-explicit system (3.1) is given by Brenan & Petzold (1989). Their main result is that for (4.1) the global error of the y -component is $\mathcal{O}(h^{q+1})$, and that of the z -component is $\mathcal{O}(h^q)$ (where q denotes the stage order of the method). This result was improved by HLR89, using a different approach (local and global error are studied separately).

The Nonlinear System

We first investigate existence, uniqueness and the influence of perturbations to the solution of the nonlinear system (4.1). In order to simplify the notation we write (η, ζ) for (y_n, z_n) , which we assume h -dependent, and we suppress the index n in Y_{ni} , etc. The nonlinear system then reads

$$\left. \begin{aligned} Y_i &= \eta + h \sum_{j=1}^s a_{ij} f(Y_j, Z_j) \\ 0 &= g(Y_i) \end{aligned} \right\} \quad i = 1, \dots, s \quad (4.2)$$

Once a solution to (4.2) is known, we can compute ℓ_{ni} from (4.1c) (whenever (a_{ij}) is an invertible matrix) and then y_{n+1}, z_{n+1} from (4.1a).

Theorem 4.1 (HLR89, p. 31). *Suppose that (η, ζ) satisfy*

$$g(\eta) = \mathcal{O}(h^2), \quad g_y(\eta) f(\eta, \zeta) = \mathcal{O}(h) \quad (4.3)$$

and that (3.2) holds in a neighbourhood of (η, ζ) . If the Runge-Kutta matrix (a_{ij}) is invertible, then the nonlinear system (4.2) possesses for $h \leq h_0$ a locally unique solution which satisfies

$$Y_i - \eta = \mathcal{O}(h), \quad Z_i - \zeta = \mathcal{O}(h). \quad (4.4)$$

Remark. Condition (4.3) expresses the fact that (η, ζ) is close to consistent initial values. We also see from (4.2) that the solution (Y_i, Z_i) does not depend on ζ . The value of ζ in (4.3) only specifies the solution branch of $g_y(y)f(y, z) = 0$ to which the numerical solution is close.

The *proof* of Theorem 4.1 for the implicit Euler method was given in Sect. VII.3 (proof of Theorem 3.1). If we replace (3.14) by

$$Y_i - \eta(h) - h \sum_{j=1}^s a_{ij} f(Y_j, Z_j) = 0 \quad (4.5a)$$

$$\int_0^1 g_y \left(\eta(h) + \tau(Y_i - \eta(h)) \right) d\tau \cdot \sum_{j=1}^s a_{ij} f(Y_j, Z_j) + \frac{1}{h} g(\eta(h)) = 0 \quad (4.5b)$$

it extends in a straightforward manner to general Runge-Kutta methods. \square

Influence of Perturbations. Besides (4.2) we also consider the perturbed system

$$\left. \begin{aligned} \hat{Y}_i &= \hat{\eta} + h \sum_{j=1}^s a_{ij} f(\hat{Y}_j, \hat{Z}_j) + h \delta_i \\ 0 &= g(\hat{Y}_i) + \theta_i \end{aligned} \right\} \quad i = 1, \dots, s \quad (4.6)$$

and we investigate the influence of the perturbations δ_i and θ_i on the numerical solution.

Theorem 4.2 (HLR89, p. 33). *Let Y_i, Z_i be a solution of (4.2) and consider perturbed values \hat{Y}_i, \hat{Z}_i satisfying (4.6). In addition to the assumptions of Theorem 4.1 suppose that*

$$\hat{\eta} - \eta = \mathcal{O}(h^2), \quad \hat{Z}_i - \zeta = \mathcal{O}(h), \quad \delta_i = \mathcal{O}(h), \quad \theta_i = \mathcal{O}(h^2). \quad (4.7)$$

Then we have for $h \leq h_0$ the estimates

$$\|\hat{Y}_i - Y_i\| \leq C \left(\|\hat{\eta} - \eta\| + h\|\delta\| + \|\theta\| \right) \quad (4.8a)$$

$$\|\hat{Z}_i - Z_i\| \leq \frac{C}{h} \left(\|g_y(\eta)(\hat{\eta} - \eta)\| + h\|\hat{\eta} - \eta\| + h\|\delta\| + \|\theta\| \right) \quad (4.8b)$$

where $\|\delta\| = \max_i \|\delta_i\|$ and $\|\theta\| = \max_i \|\theta_i\|$. If the initial values satisfy $g(\eta) = 0$ and $g(\hat{\eta}) = 0$, then we have the stronger estimate

$$\|\hat{Z}_i - Z_i\| \leq \frac{C}{h} \left(h\|\hat{\eta} - \eta\| + h\|\delta\| + \|\theta\| \right). \quad (4.9)$$

The constant C in (4.8) and (4.9) depends only on bounds for certain derivatives of f and g , but not on the constants implied by the $\mathcal{O}(\dots)$ terms in (4.3) and (4.7).

Proof. The estimates (4.8) are obtained by extending the proof of Theorem 3.2. When both initial values, η and $\hat{\eta}$, lie on the manifold $g(y) = 0$, we have by Taylor expansion $0 = g(\hat{\eta}) - g(\eta) = g_y(\eta)(\hat{\eta} - \eta) + \mathcal{O}(\|\hat{\eta} - \eta\|^2)$. In this situation the term $g_y(\eta)(\hat{\eta} - \eta)$ in (4.8b) is of size $\mathcal{O}(h^2\|\hat{\eta} - \eta\|)$ and may be neglected. \square

Estimation of the Local Error

We begin by defining two projections which will be important for the study of local errors for index 2 problems (3.1).

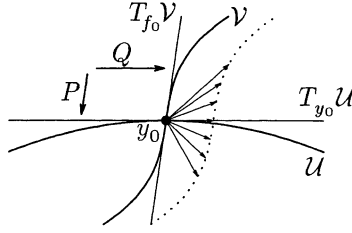
Definition 4.3. For given y_0, z_0 for which $(g_y f_z)(y_0, z_0)$ is invertible we define the projections

$$Q = (f_z(g_y f_z)^{-1} g_y)(y_0, z_0), \quad P = I - Q. \quad (4.10)$$

Geometric interpretation. Let \mathcal{U} be the manifold defined by $\mathcal{U} = \{y; g(y) = 0\}$ and let $T_{y_0}\mathcal{U} = \ker(g_y(y_0))$ be the tangent space at a point $y_0 \in \mathcal{U}$. Further let $\mathcal{V} = \{f(y_0, z); z \text{ arbitrary}\}$ and let $T_{f_0}\mathcal{V} = \text{Im}(f_z(y_0, z_0))$ be its tangent space at $f_0 = f(y_0, z_0)$. Here, z_0 is the value for which $f(y_0, z_0)$ lies in $T_{y_0}\mathcal{U}$ (i.e., for which the condition $g_y(y_0)f(y_0, z_0) = 0$ is satisfied (see 1.14c)). By considering the arrows $f(y_0, z)$ with varying z (see Fig. 4.1), the space $T_{f_0}\mathcal{V}$ can be interpreted as the directions in which the control variables z bring the solution to the manifold \mathcal{U} . By (3.2) these two spaces are transversal and their direct sum generates the y -space. It follows from (4.10) that P projects onto $T_{y_0}\mathcal{U}$ parallel to $T_{f_0}\mathcal{V}$ and Q projects onto $T_{f_0}\mathcal{V}$ parallel to $T_{y_0}\mathcal{U}$.

Consider now initial values $y_0 = y(x)$, $z_0 = z(x)$ on the exact solution and denote by y_1, z_1 the numerical solution of the Runge-Kutta method (4.1). The local error

$$\delta y_h(x) = y_1 - y(x+h), \quad \delta z_h(x) = z_1 - z(x+h) \quad (4.11)$$

Fig. 4.1. Projections P and Q

can be estimated as follows:

Lemma 4.4 (HLR89, p. 34). *Suppose that a Runge-Kutta method with invertible coefficient matrix (a_{ij}) satisfies the assumptions $B(p)$ and $C(q)$ of Sect. IV.5 with $p \geq q$. Then we have*

$$\begin{aligned} \delta y_h(x) &= \mathcal{O}(h^{q+1}), & P(x)\delta y_h(x) &= \mathcal{O}(h^{\min(p+1, q+2)}) \\ \delta z_h(x) &= \mathcal{O}(h^q), \end{aligned} \quad (4.12)$$

where $P(x)$ is the projection (4.10) evaluated at $(y(x), z(x))$. If, in addition, the Runge-Kutta method is stiffly accurate (i.e., satisfies $a_{si} = b_i$ for all i), then

$$\delta y_h(x) = \mathcal{O}(h^{\min(p+1, q+2)}). \quad (4.13)$$

Proof. The exact solution values $\hat{\eta} = y(x)$, $\hat{Y}_i = y(x + c_i h)$, $\hat{Z}_i = z(x + c_i h)$ satisfy (4.6) with $\theta_i = 0$ and

$$\delta_i = \frac{h^q}{q!} y^{(q+1)}(x) \left(\frac{c_i^{q+1}}{q+1} - \sum_{j=1}^s a_{ij} c_j^q \right) + \mathcal{O}(h^{q+1}).$$

The difference to the numerical solution ((4.2) with $\eta = y(x)$) can thus be estimated with Theorem 4.2, yielding

$$Y_i - y(x + c_i h) = \mathcal{O}(h^{q+1}), \quad Z_i - z(x + c_i h) = \mathcal{O}(h^q). \quad (4.14)$$

Since the quadrature formula $\{b_i, c_i\}$ is of order p , we have

$$y(x+h) - y(x) - h \sum_{i=1}^s b_i f(y(x + c_i h), z(x + c_i h)) = \mathcal{O}(h^{p+1}).$$

Subtracting this formula from (4.1a) we get

$$y_1 - y(x+h) = h f_z(y(x), z(x)) \sum_{i=1}^s b_i (Z_i - z(x + c_i h)) + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+2}).$$

Because of $P(x)f_z(y(x), z(x)) \equiv 0$, this proves (4.12) for the y -component. The

estimate for the z -component follows from (see (1.28))

$$z_1 - z(x+h) = \sum_{i,j=1}^s b_i \omega_{ij} (Z_j - z(x+c_j h)) + \mathcal{O}(h^{q+1})$$

and (4.14).

Under the assumption $a_{si} = b_i$ (for all i) we have $g(y_1) = 0$ so that by Taylor expansion

$$0 = g(y_1) - g(y(x+h)) = g_y(y(x))\delta y_h(x) + \mathcal{O}(h\|\delta y_h(x)\|). \quad (4.15)$$

This implies that $Q(x)\delta y_h(x) = \mathcal{O}(h\|\delta y_h(x)\|)$, and (4.13) is a consequence of (4.12) and (4.10). \square

For some important Runge-Kutta methods (such as Radau IIA and Lobatto IIIC) the estimates of Lemma 4.4 are not optimal. Sharp estimates will be given in Theorem 4.9 for collocation methods and in Sect. VII.5 for general Runge-Kutta methods.

Convergence for the y -Component

The numerical solution $\{y_n\}$, defined by (4.1), does not depend on $\{z_n\}$. Consequently, the convergence for the y -component can be treated independently of estimates for the z -component.

Theorem 4.5 (HLR89, p. 36). *Suppose that (3.2) holds in a neighbourhood of the solution $(y(x), z(x))$ of (3.1) and that the initial values are consistent. Suppose further that the Runge-Kutta matrix (a_{ij}) is invertible, that $|R(\infty)| < 1$ (see (VI.1.11e)) and that the local error satisfies*

$$\delta y_h(x) = \mathcal{O}(h^r), \quad P(x)\delta y_h(x) = \mathcal{O}(h^{r+1}) \quad (4.16)$$

with $P(x)$ as in Lemma 4.4. Then the method (4.1) is convergent of order r , i.e.,

$$y_n - y(x_n) = \mathcal{O}(h^r) \quad \text{for } x_n - x_0 = nh \leq \text{Const.}$$

If in addition $\delta y_h(x) = \mathcal{O}(h^{r+1})$, then $g(y_n) = \mathcal{O}(h^{r+1})$.

Proof. A complete proof of this result is given in (HLR89, pp. 36-39). We restrict our presentation to stiffly accurate Runge-Kutta methods (i.e., $a_{si} = b_i$ for all i). This considerably simplifies several parts of the proof, and nevertheless covers many important Runge-Kutta methods (such as Radau IIA, Lobatto IIIC and the SDIRK method (IV.6.16)). The assumption $a_{si} = b_i$ (for all i) implies that $g(y_n) = 0$ for all n and, as a consequence of (4.15) and (4.16), that

$$\delta y_h(x) = \mathcal{O}(h^{r+1}). \quad (4.17)$$

The following proof is similar to that of Theorem 3.5 and uses, once again, Lady Windermere's Fan of Fig. II.3.2.

In addition to the numerical solution $\{y_n, z_n\}$, also denoted by $\{y_n^0, z_n^0\}$, we consider the Runge-Kutta solutions $\{y_n^\ell, z_n^\ell\}$ with initial values $y_\ell^\ell = y(x_\ell)$, $z_\ell^\ell = z(x_\ell)$ on the exact solution. We first estimate $y_n^\ell - y_n^{\ell+1}$ for $n \geq \ell + 1$ in terms of the local error $\delta y_h(x_\ell) = y_{\ell+1}^\ell - y_{\ell+1}^{\ell+1}$. In order to simplify the notation we denote two neighbouring Runge-Kutta solutions by $\{\tilde{y}_n\}$, $\{\hat{y}_n\}$ and their difference by $\Delta y_n = \tilde{y}_n - \hat{y}_n$. We suppose for the moment that

$$\|\hat{y}_n - y(x_n)\| \leq C_0 h, \quad \|\Delta y_n\| \leq C_1 h^2 \quad (4.18)$$

(this will be justified below). Theorem 4.2 with $\delta_i = 0$ and $\theta_i = 0$ then yields

$$\|\tilde{Y}_{ni} - \hat{Y}_{ni}\| \leq C \|\Delta y_n\|, \quad \|\tilde{Z}_{ni} - \hat{Z}_{ni}\| \leq C \|\Delta y_n\| \quad (4.19)$$

where C is some constant independent of C_0 and C_1 . A Lipschitz condition for $f(y, z)$ implies that

$$\|\Delta y_{n+1}\| \leq \|\Delta y_n\| + h \sum_{i=1}^s |b_i| \left(L_1 \|\tilde{Y}_{ni} - \hat{Y}_{ni}\| + L_2 \|\tilde{Z}_{ni} - \hat{Z}_{ni}\| \right).$$

Inserting (4.19) we get $\|\Delta y_{n+1}\| \leq (1 + hL) \|\Delta y_n\|$ and hence also

$$\|\Delta y_n\| \leq C_2 \|\Delta y_0\| \quad \text{for } nh \leq \text{Const.} \quad (4.20)$$

For our situation in Lady Windermere's Fan the use of (4.17) yields

$$\|y_n^\ell - y_n^{\ell+1}\| \leq C_2 \|\delta y_h(x_\ell)\| \leq C_3 h^{r+1} \quad \text{for } n \geq \ell + 1 \text{ and } nh \leq \text{Const.}$$

Summing up we obtain the desired estimate

$$\|y_n - y(x_n)\| \leq \sum_{\ell=0}^{n-1} \|y_n^\ell - y_n^{\ell+1}\| \leq C_4 h^r \quad \text{for } nh \leq \text{Const.}$$

Since C_3 and C_4 do not depend on C_0 or C_1 (if h is sufficiently small), the assumption (4.18) is justified by induction on n provided the constants C_0, C_1 are chosen sufficiently large. \square

Convergence for the z -Component

Theorem 4.6 (HLR89, p. 40). *Consider the index 2 problem (3.1)–(3.2) with consistent initial values and assume that the Runge-Kutta matrix (a_{ij}) is invertible and $|R(\infty)| < 1$. If the global error of the y -component is $\mathcal{O}(h^r)$, $g(y_n) = \mathcal{O}(h^{r+1})$ and the local error of the z -component is $\mathcal{O}(h^r)$, then we have for the global error*

$$z_n - z(x_n) = \mathcal{O}(h^r) \quad \text{for } x_n - x_0 = nh \leq \text{Const.}$$

Remark. If, in addition to the invertibility of (a_{ij}) and $|R(\infty)| < 1$, the conditions $B(q)$ and $C(q)$ are satisfied then we have $z_n - z(x_n) = \mathcal{O}(h^q)$ (see Lemma 4.4).

Proof. We write the global error as

$$z_{n+1} - z(x_{n+1}) = z_{n+1} - \widehat{z}_{n+1} + \delta z_h(x_n) \quad (4.21)$$

where $(\widehat{y}_{n+1}, \widehat{z}_{n+1})$ denotes the numerical solution obtained from the starting values $(y(x_n), z(x_n))$ and $\delta z_h(x_n)$ is the local error. From (VI.1.11d) we have

$$z_{n+1} - \widehat{z}_{n+1} = R(\infty)(z_n - z(x_n)) + \sum_{i,j=1}^s b_i \omega_{ij}(Z_{nj} - \widehat{Z}_{nj}). \quad (4.22)$$

The assumption $g(y_n) = \mathcal{O}(h^{r+1})$ implies that $g_y(y_n)(y_n - y(x_n)) = \mathcal{O}(h^{r+1})$ and, together with $y_n - y(x_n) = \mathcal{O}(h^r)$, it follows from Theorem 4.2 that $Z_{nj} - \widehat{Z}_{nj} = \mathcal{O}(h^r)$. Inserting (4.22) into (4.21) we obtain

$$z_{n+1} - z(x_{n+1}) = R(\infty)(z_n - z(x_n)) + \mathcal{O}(h^r),$$

which proves the statement. \square

Collocation Methods

An important subclass of implicit Runge-Kutta methods are the collocation methods as introduced in Sect. II.7. For the index 2 problem (3.1) they can be defined as follows.

Definition 4.7. Let c_1, \dots, c_s be s distinct real numbers and denote by $u(x), v(x)$ the polynomials of degree s (*collocation polynomials*) which satisfy

$$u(x_0) = y_0, \quad v(x_0) = z_0 \quad (4.23a)$$

$$\left. \begin{aligned} u'(x_0 + c_i h) &= f(u(x_0 + c_i h), v(x_0 + c_i h)) \\ 0 &= g(u(x_0 + c_i h)) \end{aligned} \right\} \quad i = 1, \dots, s. \quad (4.23b)$$

Then, the numerical solution is given by

$$y_1 = u(x_0 + h), \quad z_1 = v(x_0 + h). \quad (4.23c)$$

A straightforward extension of Theorems II.7.7 and II.7.8 to index 2 problems shows that (4.23) is equivalent to the s -stage Runge-Kutta method (4.1) whose coefficients are defined by $B(s)$ and $C(s)$ (see Sect. IV.5 for their definition). This equivalence allows us to deduce from Theorem 4.1 the existence and local uniqueness of the collocation polynomials provided that the corresponding Runge-Kutta matrix is invertible. Hence we assume in the sequel that $c_i \neq 0$ for all i . The case of a singular Runge-Kutta matrix is considered in Exercises 2 and 3.

The quality of $u(x), v(x)$ as approximations to $y(x), z(x)$ is described by the next theorem, which extends Theorem II.7.10.

Theorem 4.8. Consider a collocation method (4.23) with all $c_i \neq 0$. Then we have for $k = 0, 1, \dots, s$ and $x \in [x_0, x_0 + h]$

$$\begin{aligned}\|u^{(k)}(x) - y^{(k)}(x)\| &\leq C h^{s+1-k}, \\ \|v^{(k)}(x) - z^{(k)}(x)\| &\leq C h^{s-k}.\end{aligned}$$

Proof. We exploit the fact that $u(x_0 + c_i h) = Y_i$, $v(x_0 + c_i h) = Z_i$ are the internal stages of the Runge-Kutta method (4.1). Consequently the collocation polynomials can be written as

$$u(x_0 + th) = y_0 \ell_0(t) + \sum_{i=1}^s Y_i \ell_i(t) \quad (4.24a)$$

$$v(x_0 + th) = z_0 \ell_0(t) + \sum_{i=1}^s Z_i \ell_i(t) \quad (4.24b)$$

where the $\ell_i(t)$ are the Lagrange polynomials defined by

$$\ell_0(t) = \prod_{j=1}^s \frac{(t - c_j)}{(-c_j)}, \quad \ell_i(t) = \frac{t}{c_i} \prod_{\substack{j=1 \\ j \neq i}}^s \frac{(t - c_j)}{(c_i - c_j)}.$$

Familiar estimates of the interpolation error imply that the exact solution $y(x)$ satisfies

$$y(x_0 + th) = y_0 \ell_0(t) + \sum_{i=1}^s y(x_0 + c_i h) \ell_i(t) + \mathcal{O}(h^{s+1}). \quad (4.25)$$

The factor h^{s+1} in the interpolation error comes from the $(s+1)$ -th derivative of $y(x_0 + th)$ with respect to t . Obviously, the interpolation error is differentiable as often as the function $y(x)$. If we differentiate (4.25) k times, then by Rolle's theorem, the difference

$$h^k y^{(k)}(x_0 + th) - \left(y_0 \ell_0^{(k)}(t) + \sum_{i=1}^s y(x_0 + c_i h) \ell_i^{(k)}(t) \right) \quad (4.25')$$

vanishes at least at $s+1-k$ points. Hence, the polynomial enclosed in brackets in (4.25') can be interpreted as an interpolation polynomial of degree $s-k$ for the function $h^k y^{(k)}(x_0 + th)$. Its error is thus again of size $\mathcal{O}(h^{s+1})$. Subtracting (4.25) from (4.24a) and differentiating k times thus yields

$$h^k (u^{(k)}(x_0 + th) - y^{(k)}(x_0 + th)) = \sum_{i=1}^s (Y_i - y(x_0 + c_i h)) \ell_i^{(k)}(t) + \mathcal{O}(h^{s+1})$$

and a similar formula for the z -component. The conclusion now follows from (4.14) with $q = s$. \square

Superconvergence of Collocation Methods

It is now natural to ask whether superconvergence takes place at $x_0 + h$ (as for ordinary differential equations; see Theorem II.7.9). The answer is affirmative, if the method is stiffly accurate, i.e., if $c_s = 1$.

Theorem 4.9. *If $c_i \neq 0$ for all i and $c_s = 1$, then the y -component of the local error of the collocation method (4.23) satisfies*

$$y_1 - y(x_0 + h) = \mathcal{O}(h^{p+1}),$$

where p is the order of the underlying quadrature formula.

Proof. We insert the collocation polynomials into the differential-algebraic problem and define the defect by

$$u'(x) = f(u(x), v(x)) + \delta(x) \quad (4.26a)$$

$$0 = g(u(x)) + \theta(x). \quad (4.26b)$$

By Definition 4.7 we have

$$\delta(x_0 + c_i h) = 0, \quad \theta(x_0) = 0, \quad \theta(x_0 + c_i h) = 0. \quad (4.27)$$

We next differentiate (4.26b) with respect to x and use (4.26a):

$$0 = g_y(u(x))(f(u(x), v(x)) + \delta(x)) + \theta'(x). \quad (4.28)$$

This motivates the use of the equation

$$0 = g_y(u)(f(u, v) + \delta(x)) + \theta'(x) \quad (4.29)$$

for arbitrary (u, v) in a neighbourhood of the solution of (3.1). Because of (3.2) we can extract v from (4.29) so that (4.29) can be written as

$$v = G(u, \delta(x), \theta'(x)). \quad (4.30)$$

Inserting into (4.26a) and into (3.1) this yields

$$u'(x) = f(u(x), G(u(x), \delta(x), \theta'(x))) + \delta(x) \quad (4.31a)$$

$$y'(x) = f(y(x), G(y(x), 0, 0)). \quad (4.31b)$$

In order to compute $u(x) - y(x)$ we now apply the nonlinear variation-of-constants formula (Theorem I.14.5). This requires the computation of the defect of $u(x)$ inserted into (4.31b)

$$\begin{aligned} & u'(x) - f(u(x), G(u(x), 0, 0)) \\ &= f(u(x), G(u(x), \delta(x), \theta'(x))) + \delta(x) - f(u(x), G(u(x), 0, 0)) \\ &= \Phi(x, 1) - \Phi(x, 0) + \delta(x) \end{aligned} \quad (4.32)$$

where

$$\Phi(x, \tau) = f\left(u(x), G(u(x), \tau \cdot \delta(x), \tau \cdot \theta'(x))\right).$$

Then the formula $\Phi(x, 1) - \Phi(x, 0) = \int_0^1 \partial\Phi/\partial\tau(x, \tau) d\tau$ shows that the defect (4.32) can be written as

$$Q_1(x)\delta(x) + Q_2(x)\theta'(x). \quad (4.32')$$

We now insert this into Eq. (I.14.18) and obtain

$$\begin{aligned} u(x) - y(x) &= \int_{x_0}^x \text{resolvent}(x, t) \cdot \text{defect}(t) dt \\ &= \int_{x_0}^x \left(S_1(x, t)\delta(t) + S_2(x, t)\theta'(t) \right) dt. \end{aligned}$$

Integrating the second term by parts we get (since $\theta(x_0) = 0$)

$$\begin{aligned} y_1 - y(x_0 + h) &= \int_{x_0}^{x_0+h} \left(S_1(x_0 + h, t)\delta(t) - \frac{\partial S_2}{\partial t}(x_0 + h, t)\theta(t) \right) dt \\ &\quad + S_2(x_0 + h, x_0 + h)\theta(x_0 + h). \end{aligned} \quad (4.33)$$

The assumption $c_s = 1$ implies that $\theta(x_0 + h) = 0$ so that the last expression in (4.33) vanishes. The main idea is now to integrate the expression in (4.33) with the quadrature formula $\{b_i, c_i\}$ (see also the proof of Theorem II.7.9). With the abbreviation

$$\sigma(t) = S_1(x_0 + h, t)\delta(t) - \frac{\partial S_2}{\partial t}(x_0 + h, t)\theta(t) \quad (4.34)$$

this gives

$$y_1 - y(x_0 + h) = \int_{x_0}^{x_0+h} \sigma(t) dt = h \sum_{i=1}^s b_i \sigma(x_0 + c_i h) + \text{err}(\sigma). \quad (4.35)$$

Because of (4.27) we have $\sigma(x_0 + c_i h) = 0$ for all i and the quadrature error is estimated by

$$\|\text{err}(\sigma)\| \leq Ch^{p+1} \max_{t \in [x_0, x_0+h]} \|\sigma^{(p)}(t)\|. \quad (4.36)$$

The p -th derivative of $\sigma(t)$ contains derivatives of f, g and of $\delta(x), \theta(x)$. By Theorem 4.8 they are uniformly bounded for $h \leq h_0$. Hence $y_1 - y(x_0 + h) = \text{err}(\sigma) = \mathcal{O}(h^{p+1})$, proving the theorem. \square

Projected Runge-Kutta Methods

For collocation methods which are not stiffly accurate it is possible to prove super-convergence (as in Theorem 4.9) if the method is combined with a certain projection. We start with a more careful study of the local error of the y -component in (4.33).

Lemma 4.10. *If $c_i \neq 0$ for all i , then the y -component of the local error of the collocation method (4.23) satisfies*

$$y_1 - y(x_0 + h) = -\left(f_z(g_y f_z)^{-1}\right)(y(x_0 + h), z(x_0 + h))\theta(x_0 + h) + \mathcal{O}(h^{p+1}) \quad (4.37)$$

where θ is the defect given by (4.26b) and p is the order of the underlying quadrature formula.

Proof. The above proof of Theorem 4.9 (see Eq. (4.33)) shows that the local error satisfies

$$y_1 - y(x_0 + h) = S_2(x_0 + h, x_0 + h)\theta(x_0 + h) + \mathcal{O}(h^{p+1}).$$

Hence, we only have to compute $S_2(x, x)$. Since any resolvent equals the identity matrix if both of its arguments are equal, it follows from the definition of $S_2(x, t)$ and from (4.32') that

$$S_2(x, x) = \int_0^1 f_z(u(x), G(u(x), \tau\delta(x), \tau\theta'(x))) \frac{\partial G}{\partial \theta'}(u(x), \tau\delta(x), \tau\theta'(x)) d\tau.$$

Differentiating (4.29) with respect to θ' gives

$$\frac{\partial G}{\partial \theta'} = \frac{\partial v}{\partial \theta'} = -(g_y f_z)^{-1}(u, v).$$

Furthermore, it follows from (4.27) that $\delta(x) = \mathcal{O}(h^s)$ and $\theta'(x) = \mathcal{O}(h^s)$ for $x = x_0 + h$. Using $u(x) - y(x) = \mathcal{O}(h^{s+1})$ (from Theorem 4.8) we thus obtain for $x = x_0 + h$

$$S_2(x, x) = \left(f_z(g_y f_z)^{-1}\right)(y(x), z(x)) + \mathcal{O}(h^s).$$

The statement now follows from $p \leq 2s$ and from $\theta(x_0 + h) = \mathcal{O}(h^{s+1})$. □

The geometric interpretation of Lemma 4.10 is as follows: if we split the local error $\delta y_h(x_0)$ according to the projections of Fig. 4.1 then the component $Q(x_0 + h)\delta y_h(x_0)$ is of size $\mathcal{O}(h^{s+1})$, whereas the component $P(x_0 + h)\delta y_h(x_0)$ is $\mathcal{O}(h^{p+1})$. This suggests to project after every step the numerical solution of a Runge-Kutta method onto the manifold $g(y) = 0$ with the help of the projection operator $P(x_0 + h)$ as follows:

Definition 4.11 (Ascher & Petzold 1991). Let y_1, z_1 be the numerical solution of an implicit Runge-Kutta method (4.1) and define \hat{y}_1, λ as the solution of the system

$$\begin{aligned}\hat{y}_1 &= y_1 + f_z(\hat{y}_1, z_1)\lambda \\ 0 &= g(\hat{y}_1).\end{aligned}\tag{4.38}$$

If the value \hat{y}_1 (and z_1) is used for the step by step integration of (3.1), then we call this procedure *projected Runge-Kutta method*.

Remarks. 1) If $g(y_1)$ is sufficiently small, then the nonlinear system (4.38) possesses a locally unique solution. A Newton-type iteration with starting values $\hat{y}_1^{(0)} = y_1, \lambda^{(0)} = 0$ will converge to this solution. This follows at once from the theorem of Newton-Kantorovich (Ortega & Rheinboldt 1970) because the Jacobian of (4.38) evaluated at the starting values

$$\begin{pmatrix} I & -f_z(y_1, z_1) \\ g_y(y_1) & 0 \end{pmatrix}$$

has a bounded inverse by (3.2).

2) For stiffly accurate Runge-Kutta methods (i.e., if $a_{si} = b_i$ for all i) the projected and unprojected Runge-Kutta methods coincide.

3) The proof of the next theorem shows that the argument in $f_z(\hat{y}_1, z_1)$ may be replaced by some other approximation to $y(x_0 + h), z(x_0 + h)$ whose error is at most $\mathcal{O}(h^s)$.

The following theorem proves superconvergence for projected collocation methods (also if the corresponding Runge-Kutta method is not stiffly accurate). Superconvergence results for general Runge-Kutta methods are given in Sect. VI.8.

Theorem 4.12 (Ascher & Petzold 1991). *If $c_i \neq 0$ for all i , then the y -component of the local error of the projected collocation method (4.23), (4.38) satisfies*

$$\hat{y}_1 - y(x_0 + h) = \mathcal{O}(h^{p+1})$$

where p is the order of the underlying quadrature formula.

Proof. We write $\hat{e}_1 = \hat{y}_1 - y(x_0 + h), e_1 = y_1 - y(x_0 + h)$ for the local errors and denote the projections of Definition 4.3 by

$$Q = (f_z(g_y f_z)^{-1} g_y)(\hat{y}_1, z_1), \quad P = I - Q.$$

The idea is to split \hat{e}_1 according to

$$\hat{e}_1 = P \hat{e}_1 + Q \hat{e}_1\tag{4.39}$$

and to estimate both components separately. The first formula of (4.38) together with (4.37) and $\theta(x_0 + h) = \mathcal{O}(h^{s+1})$ imply that

$$P \hat{e}_1 = P e_1 = \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{s+1} \|\hat{e}_1\|).\tag{4.40}$$

Further we have $0 = g(\widehat{y}_1) - g(y(x_0 + h)) = g_y(\widehat{y}_1)\widehat{e}_1 + \mathcal{O}(\|\widehat{e}_1\|^2)$, implying

$$Q\widehat{e}_1 = \mathcal{O}(\|\widehat{e}_1\|^2). \tag{4.41}$$

Formulas (4.40) and (4.41) inserted into (4.39) give

$$\widehat{e}_1 = \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{s+1}\|\widehat{e}_1\|) + \mathcal{O}(\|\widehat{e}_1\|^2)$$

and the statement of the theorem is an immediate consequence. □

Global convergence of order $\mathcal{O}(h^p)$ of the projected collocation methods is obtained exactly as in the proof of Theorem 4.5. We observe that the numerical solution always remains on the manifold $g(y) = 0$ so that the estimate (4.9) applies.

Summary of Convergence Results

Table 4.1 collects the optimal error estimates for some important Runge-Kutta methods when applied to the index 2 problem (3.1)–(3.2). The local error estimates can be verified as follows: Gauss, Radau IA and SDIRK by Lemma 4.4, Radau IIA by Theorem 4.9, Lobatto IIIC by Theorem 5.10 below and Lobatto IIIA with the help of Exercise 4. For the projected methods the estimates follow from Theorem 4.12 and the considerations of Sect. VII.5. Because there are several ways of defining the z -component of the numerical solution, we do not present their convergence behaviour. The global convergence result follows from Theorems 4.5 and 4.6 for the Radau IA, Radau IIA, Lobatto IIIC and SDIRK methods. The remaining methods (Gauss and Lobatto IIIA) require some more effort because their stability function only satisfies $|R(\infty)| = 1$. For a detailed discussion of these methods we refer to HLR89 and Jay (1993).

Table 4.1. Error estimates for the index 2 problem (3.1)–(3.2)

Method	stages	local error		global error	
		y	z	y	z
Gauss	$\begin{cases} s & \text{odd} \\ s & \text{even} \end{cases}$	h^{s+1}	h^s	$\begin{cases} h^{s+1} \\ h^s \end{cases}$	$\begin{cases} h^{s-1} \\ h^{s-2} \end{cases}$
projected Gauss	s	h^{2s+1}		h^{2s}	
Radau IA	s	h^s	h^{s-1}	h^s	h^{s-1}
projected Radau IA	s	h^{2s-1}		h^{2s-2}	
Radau IIA	s	h^{2s}	h^s	h^{2s-1}	h^s
Lobatto IIIA	$\begin{cases} s & \text{odd} \\ s & \text{even} \end{cases}$	h^{2s-1}	h^s	h^{2s-2}	$\begin{cases} h^{s-1} \\ h^s \end{cases}$
Lobatto IIIC	s	h^{2s-1}	h^{s-1}	h^{2s-2}	h^{s-1}
SDIRK (IV.6.16)	5	h^3	h^1	h^2	h^1
SDIRK (IV.6.18)	3	h^2	h^1	h^2	h^1

Exercises

1. Consider the index 2 problem $y' = f(y, z)$, $0 = g(y)$. Put $z = u'$, $v = (y, u)^T$ so that the problem becomes

$$F(v', v) = \begin{pmatrix} y' - f(y, u') \\ g(y) \end{pmatrix} = 0.$$

Prove that the matrix pencil $F_v + \lambda F_{v'}$ is of index 1 whenever $(g_y f_z)^{-1}$ exists.

Hint. Consider the transformation

$$\begin{pmatrix} I & a \\ 0 & I \end{pmatrix} (F_v + \lambda F_{v'}) \begin{pmatrix} I & b \\ 0 & I \end{pmatrix} \quad (4.42)$$

where $a = f_y f_z (g_y f_z)^{-1}$ and $b = f_z$ are chosen such that the upper right block in (4.42) vanishes.

2. Consider Runge-Kutta methods whose coefficients satisfy:

$$a_{1i} = 0 \text{ for all } i \text{ and } (a_{ij})_{i,j \geq 2} \text{ is invertible.}$$

(Examples are collocation methods with $c_1 = 0$, such as Lobatto IIIA).

If $g(\eta) = 0$ then the nonlinear system (4.2) has a locally unique solution which satisfies $Y_1 = \eta$, $Z_1 = \zeta$.

3. Let $c_1 = 0$, c_2, \dots, c_s be s distinct real numbers. Show that there exist unique polynomials $u(x)$ and $v(x)$ ($\deg u = s$, $\deg v = s - 1$) such that (4.23a,b) holds.

Hint. Apply the ideas of the proof of Theorem II.7.7 and Exercise 2.

4. Investigate the validity of the conclusions of Theorems 4.8 and 4.9 for the situation where $c_1 = 0$.
5. (Computation of the algebraic variable z by *piecewise discontinuous interpolation*, see Ascher (1989)). Modify the definition of z_{n+1} in the Runge-Kutta method (4.1) as follows: let $v(x)$ be the polynomial of degree $s - 1$ satisfying $v(x_n + c_i h) = Z_{ni}$ for all i , then define $z_{n+1} = v(x_n + h)$. In the case of *collocation* methods (4.23) this definition removes the condition $v(x_0) = z_0$ while lowering the degree of $v(x)$ by 1.

a) Verify: z_{n+1} does not depend on z_n , also if the stability function of the method does not vanish at infinity.

b) Prove that for projected collocation methods with $c_i \neq 0$ for all i we have $z_n - z(x_n) = \mathcal{O}(h^s)$.

c) For the projected Gauss methods compare this result with that of the standard approach.

6. The statement of Theorem 4.8 still holds, if one omits the condition $v(x_0) = z_0$ in Definition 4.7 and if one lets $v(x)$ be a polynomial of degree $s - 1$.

VII.5 Order Conditions for Index 2 DAE

For an application of the convergence result of the preceding section (Theorem 4.5) it is desirable to know the optimal values of r in (4.16). Comparing the Taylor expansions of the exact and numerical solutions we derive conditions for c_i, a_{ij}, b_j which are equivalent to (4.16). For collocation methods we recover the result of Theorem 4.9. For other methods (such as Lobatto IIIC) the estimates of Lemma 4.4 are substantially improved.

The theory of this section is given in HLR89 (Sect. 5). Our presentation is slightly different and is in complete analogy to the derivation of the index 1 order conditions of Sect. VI.4. The results of this section are here applied to Runge-Kutta methods only; analogous formulas for Rosenbrock methods can be found in Roche (1988). An independent investigation, conducted for the index 2 problem $f(y, z') = 0, z = g(y)$ by A. Kværnø (1990), leads to the same order conditions for Runge-Kutta methods.

Derivatives of the Exact Solution

We consider the index 2 problem

$$y' = f(y, z) \quad (5.1a)$$

$$0 = g(y) \quad (5.1b)$$

and assume consistent initial values y_0, z_0 . The first derivative of the solution $y(x)$ is given by (5.1a). Differentiating this equation we get

$$y'' = f_y(y, z)y' + f_z(y, z)z'. \quad (5.2)$$

In order to compute z' we differentiate (5.1b) twice

$$0 = g_y(y)y' \quad (5.3a)$$

$$0 = g_{yy}(y)(y', y') + g_y(y)y'' \quad (5.3b)$$

and insert (5.2) and (5.1a). This yields (omitting the obvious function arguments)

$$0 = g_{yy}(f, f) + g_y f_y f + g_y f_z z' \quad (5.4)$$

or equivalently

$$z' = (-g_y f_z)^{-1} g_{yy}(f, f) + (-g_y f_z)^{-1} g_y f_y f. \quad (5.5)$$

Here we have used the index 2 assumption (3.2), that $g_y f_z$ is invertible in a neighbourhood of the solution. We now differentiate (5.1a) and (5.5) with respect to x , and replace the appearing y' and z' by (5.1a) and (5.5). We use (for a constant vector u)

$$\begin{aligned} & \frac{d}{dx} (-g_y f_z)^{-1} u \\ &= (-g_y f_z)^{-1} \left(g_{yy} (f_z (-g_y f_z)^{-1} u, f) + g_y f_{zy} ((-g_y f_z)^{-1} u, f) \right. \\ & \quad \left. + g_y f_{zz} ((-g_y f_z)^{-1} u, (-g_y f_z)^{-1} g_{yy} (f, f) + (-g_y f_z)^{-1} g_y f_y f) \right) \end{aligned} \quad (5.6)$$

(cf. Formula (VI.4.7)) and thus obtain

$$y'' = f_y f + f_z (-g_y f_z)^{-1} g_{yy} (f, f) + f_z (-g_y f_z)^{-1} g_y f_y f \quad (5.7)$$

$$\begin{aligned} z'' &= (-g_y f_z)^{-1} g_{yyy} (f, f, f) + 3(-g_y f_z)^{-1} g_{yy} (f, f_y f) \\ & \quad + 3(-g_y f_z)^{-1} g_{yy} (f, f_z (-g_y f_z)^{-1} g_{yy} (f, f)) \\ & \quad + 3(-g_y f_z)^{-1} g_{yy} (f, f_z (-g_y f_z)^{-1} g_y f_y f) + (-g_y f_z)^{-1} g_y f_{yy} (f, f) \\ & \quad + 2(-g_y f_z)^{-1} g_y f_{yz} (f, (-g_y f_z)^{-1} g_{yy} (f, f)) \\ & \quad + 2(-g_y f_z)^{-1} g_y f_{yz} (f, (-g_y f_z)^{-1} g_y f_y f) + (-g_y f_z)^{-1} g_y f_y f_y f \\ & \quad + (-g_y f_z)^{-1} g_y f_y f_z (-g_y f_z)^{-1} g_{yy} (f, f) \\ & \quad + (-g_y f_z)^{-1} g_y f_y f_z (-g_y f_z)^{-1} g_y f_y f \\ & \quad + (-g_y f_z)^{-1} g_y f_{zz} ((-g_y f_z)^{-1} g_{yy} (f, f), (-g_y f_z)^{-1} g_{yy} (f, f)) \\ & \quad + 2(-g_y f_z)^{-1} g_y f_{zz} ((-g_y f_z)^{-1} g_{yy} (f, f), (-g_y f_z)^{-1} g_y f_y f) \\ & \quad + (-g_y f_z)^{-1} g_y f_{zz} ((-g_y f_z)^{-1} g_y f_y f, (-g_y f_z)^{-1} g_y f_y f). \end{aligned} \quad (5.8)$$

Obviously, a graphical representation of these expressions will be of great help.

Trees and Elementary Differentials

As in Sect. VI.4 we identify each occurring f with a meagre vertex, each of its derivatives with an upwards leaving branch, the expression $(-g_y f_z)^{-1} g$ with a fat vertex and the derivatives of g therein again with upwards leaving branches. The corresponding graphs for y' , z' , y'' , z'' (see Formulas (5.1a), (5.5), (5.7), (5.8)) are given in Fig. 5.1.

The derivatives of y are characterized by trees with a *meagre root* (the lowest vertex). These trees will be denoted by t or t_i , the tree consisting of the root only (for y') being τ . Derivatives of z have trees with a *fat root*. They will be denoted by u or u_i .

Definition 5.1. Let $DAT2 = DAT2_y \cup DAT2_z$ denote the set of (*differential algebraic index 2*) *trees* defined recursively by

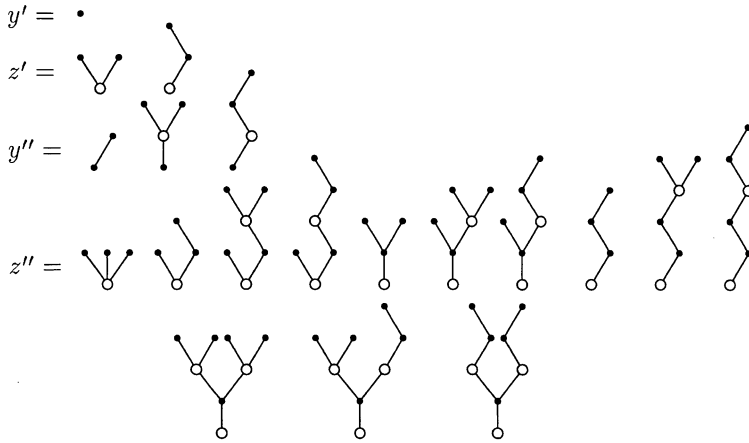


Fig. 5.1. Graphical representation of the first derivatives

- a) $\tau \in DAT2_y$,
- b) $[t_1, \dots, t_m, u_1, \dots, u_n]_y \in DAT2_y$
if $t_1, \dots, t_m \in DAT2_y$ and $u_1, \dots, u_n \in DAT2_z$;
- c) $[t_1, \dots, t_m]_z \in DAT2_z$ if $t_1, \dots, t_m \in DAT2_y$ and either $m > 1$ or
 $m = 1$ and $t_1 \neq [u]_y$ with $u \in DAT2_z$.

Definition 5.2. The *order* of a tree $t \in DAT2_y$ or $u \in DAT2_z$, denoted by $\varrho(t)$ or $\varrho(u)$, is the number of meagre vertices minus the number of fat vertices.

Definition 5.3. The *elementary differentials* $F(t)$ (or $F(u)$) corresponding to trees in $DAT2$ are defined as follows:

- a) $F(\tau) = f$,
- b) $F(t) = \frac{\partial^{m+n} f}{\partial y^m \partial z^n} \left(F(t_1), \dots, F(t_m), F(u_1), \dots, F(u_n) \right)$
if $t = [t_1, \dots, t_m, u_1, \dots, u_n]_y \in DAT2_y$,
- c) $F(u) = (-g_y f_z)^{-1} \frac{\partial^m g}{\partial y^m} \left(F(t_1), \dots, F(t_m) \right)$
if $u = [t_1, \dots, t_m]_z \in DAT2_z$.

Taylor Expansion of the Exact Solution

In order to continue the process which led to (5.7) and (5.8) we need the differentiation of elementary differentials $F(t)$ and $F(u)$. This is described by the following rules:

- i) attach to each vertex a branch with τ (derivative of f or g with respect to y and addition of the factor $y' = f$);

- ii) attach to each meagre vertex a branch with $[\tau, \tau]_z$; attach to each meagre vertex a branch with $[[\tau]_y]_z$ (this yields two trees and corresponds to the derivative of f with respect to z and to the addition of the factors $(-g_y f_z)^{-1} g_{yy}(f, f)$ and $(-g_y f_z)^{-1} g_y f_y f$ of (5.5));
- iii) split each fat vertex into two new fat vertices (one above the other) and link them via a new meagre vertex. Then four new trees are obtained as follows: attach a branch with τ to the lower of these fat vertices; attach a branch with $\tau, [\tau, \tau]_z$ or $[[\tau]_y]_z$ to the new meagre vertex (this corresponds to the derivation of $(-g_y f_z)^{-1}$ and follows at once from Eq. (5.6)).

Some of the elementary differentials in (5.8) appear more than once. In order to understand how often such an expression (or the corresponding tree) appears in the derivatives of y or z , we indicate the order of generation of the vertices as follows (see Fig. 5.2): for the trees of order 1, namely $\tau, [\tau, \tau]_z$ and $[[\tau]_y]_z$, we add the label 1 to a meagre vertex such that

each fat vertex is followed by at least one unlabelled meagre vertex. (5.9)

Each time a tree is “differentiated” according to the above rules we provide the newly attached tree (of order 1) with a new label such that (5.9) still holds. The labelling so obtained is obviously increasing along each branch.

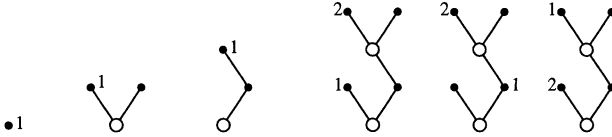


Fig. 5.2. Examples of monotonically labelled trees

Definition 5.4. A tree $t \in DAT2_y$ (or $u \in DAT2_z$), together with a monotonic labelling of $\varrho(t)$ (or $\varrho(u)$) among its meagre vertices such that (5.9) holds, is called a *monotonically labelled tree*. The sets of such monotonically labelled trees are denoted by $LDAT2_y$, $LDAT2_z$, and $LDAT2$.

Since the differentiation process of trees described above generates all elements of $LDAT2$, and each of them exactly once, and since each differentiation increases the order of the trees by one, we have the following result.

Theorem 5.5 (HLR89, p. 58). *For the exact solution of (5.1) we have:*

$$\begin{aligned}
 y^{(q)}(x_0) &= \sum_{t \in LDAT2_y, \varrho(t)=q} F(t)(y_0, z_0) = \sum_{t \in DAT2_y, \varrho(t)=q} \alpha(t) F(t)(y_0, z_0) \\
 z^{(q)}(x_0) &= \sum_{u \in LDAT2_z, \varrho(u)=q} F(u)(y_0, z_0) = \sum_{u \in DAT2_z, \varrho(u)=q} \alpha(u) F(u)(y_0, z_0).
 \end{aligned}$$

The integer coefficients $\alpha(t)$ and $\alpha(u)$ indicate the number of possible monotonic labellings of a tree. □

Derivatives of the Numerical Solution

For the problem (5.1) with consistent initial values (y_0, z_0) we write one step of a Runge-Kutta method in the form

$$y_1 = y_0 + \sum_{i=1}^s b_i k_i, \quad z_1 = z_0 + \sum_{i=1}^s b_i \ell_i \quad (5.10a)$$

where

$$k_i = h f(Y_i, Z_i), \quad 0 = g(Y_i) \quad (5.10b)$$

and

$$Y_i = y_0 + \sum_{j=1}^s a_{ij} k_j, \quad Z_i = z_0 + \sum_{j=1}^s a_{ij} \ell_j. \quad (5.10c)$$

We have replaced $h k_{ni}, h \ell_{ni}$ of Formula (4.1) by k_i, ℓ_i . This is not essential, but adjusts the derivation of the order conditions to the presentation of Sect. VI.4. Since the following derivation is very similar to the one given in Sect. VI.4, we restrict ourselves to the main ideas.

We consider $y_1, z_1, k_i, \ell_i, Y_i, Z_i$ as functions of h and compute their derivatives at $h = 0$. From (5.10a) we get

$$y_1^{(q)}(0) = \sum_{i=1}^s b_i k_i^{(q)}(0), \quad (5.11)$$

and (5.10b) yields

$$k_i^{(q)}(0) = q \left(f(Y_i, Z_i) \right)^{(q-1)} \Big|_{h=0}, \quad 0 = \left(g(Y_i) \right)^{(q)} \Big|_{h=0}. \quad (5.12)$$

The total derivatives of $f(Y_i, Z_i)$ and $g(Y_i)$ can be computed by Faà di Bruno's formula (see (VI.4.14) and (VI.4.15)). This gives

$$\left(f(Y_i, Z_i) \right)^{(q-1)} = \sum \frac{\partial^{m+n} f(Y_i, Z_i)}{\partial y^m \partial z^n} \left(Y_i^{(\mu_1)}, \dots, Y_i^{(\mu_m)}, Z_i^{(\nu_1)}, \dots, Z_i^{(\nu_n)} \right) \quad (5.13)$$

with $\mu_1 + \dots + \mu_m + \nu_1 + \dots + \nu_n = q - 1$, and

$$\left(g(Y_i) \right)^{(q)} = \sum \frac{\partial^m g(Y_i)}{\partial y^m} \left(Y_i^{(\mu_1)}, \dots, Y_i^{(\mu_m)} \right) \quad (5.14)$$

with $\mu_1 + \dots + \mu_m = q$. The summations in (5.13) and (5.14) are over sets of suitable "special labelled trees". We next insert

$$Y_i^{(\mu)} = \sum_{j=1}^s a_{ij} k_j^{(\mu)} \quad (5.15)$$

into (5.13) and (5.14) and so obtain from (5.12)

$$k_i^{(q)}(0) = q \sum \frac{\partial^{m+n} f(y_0, z_0)}{\partial y^m \partial z^n} \left(\sum_{j=1}^s a_{ij} k_j^{(\mu_1)}(0), \dots, Z_i^{(\nu_1)}(0), \dots \right) \quad (5.16)$$

and

$$0 = g_y(y_0) \sum_{j=1}^s a_{ij} k_j^{(q)}(0) + \sum_{m \geq 2} \frac{\partial^m g(y_0)}{\partial y^m} \left(\sum_{j=1}^s a_{ij} k_j^{(\mu_1)}(0), \dots \right). \quad (5.17)$$

Inserting (5.16) into the first term of (5.17) and extracting $Z_j^{(q-1)}(0)$ we get

$$\begin{aligned} & (-g_y f_z)(y_0, z_0) \sum_{j=1}^s a_{ij} Z_j^{(q-1)}(0) \\ &= \sum_{j=1}^s a_{ij} \sum_{(m,n) \neq (0,1)} g_y(y_0) \frac{\partial^{m+n} f(y_0, z_0)}{\partial y^m \partial z^n} \left(\sum_{l=1}^s a_{jl} k_l^{(\mu_1)}(0), \dots, Z_j^{(\nu_1)}(0), \dots \right) \\ & \quad + \frac{1}{q} \sum_{m \geq 2} \frac{\partial^m g(y_0)}{\partial y^m} \left(\sum_{j=1}^s a_{ij} k_j^{(\mu_1)}(0), \dots \right). \end{aligned} \quad (5.18)$$

This formula allows us to compute $Z_i^{(q-1)}$, whenever $(g_y f_z)$ and (a_{ij}) are invertible. We denote the coefficients of the inverse of (a_{ij}) by ω_{ij} , i.e.,

$$(\omega_{ij}) = (a_{ij})^{-1}. \quad (5.19)$$

The following result then follows by induction on q from (5.16) and (5.18).

Theorem 5.6 (HLR89). *The derivatives of k_i and Z_i satisfy*

$$\begin{aligned} k_i^{(q)}(0) &= \sum_{t \in \text{LDAT}_{2y}, \varrho(t)=q} \gamma(t) \Phi_i(t) F(t)(y_0, z_0) \\ Z_i^{(q)}(0) &= \sum_{u \in \text{LDAT}_{2z}, \varrho(u)=q} \gamma(u) \Phi_i(u) F(u)(y_0, z_0), \end{aligned}$$

where the coefficients $\Phi_i(t)$ and $\Phi_i(u)$ are given by $\Phi_i(\tau) = 1$ and

$$\begin{aligned} \Phi_i(t) &= \sum_{\mu_1, \dots, \mu_m} a_{i\mu_1} \cdots a_{i\mu_m} \cdot \Phi_{\mu_1}(t_1) \cdots \Phi_{\mu_m}(t_m) \Phi_i(u_1) \cdots \Phi_i(u_n) \\ & \quad \text{if } t = [t_1, \dots, t_m, u_1, \dots, u_n]_y \\ \Phi_i(u) &= \sum_{j, \mu_1, \dots, \mu_m} \omega_{ij} a_{j\mu_1} \cdots a_{j\mu_m} \cdot \Phi_{\mu_1}(t_1) \cdots \Phi_{\mu_m}(t_m) \\ & \quad \text{if } u = [t_1, \dots, t_m]_z \end{aligned}$$

and the rational coefficients $\gamma(t)$ and $\gamma(u)$ are defined by $\gamma(\tau) = 1$ and

$$\begin{aligned} \gamma(t) &= \varrho(t) \gamma(t_1) \cdots \gamma(t_m) \gamma(u_1) \cdots \gamma(u_n) & \text{if } t = [t_1, \dots, t_m, u_1, \dots, u_n]_y \\ \gamma(u) &= \frac{1}{\varrho(u) + 1} \gamma(t_1) \cdots \gamma(t_m) & \text{if } u = [t_1, \dots, t_m]_z. \end{aligned}$$

□

The derivatives of the numerical solution y_1 are now obtained from (5.11). In order to get those of z_1 , we compute ℓ_i from (5.10c) and insert it into (5.10a). This yields

$$z_1 = z_0 + \sum_{i,j=1}^s b_i \omega_{ij} (Z_j - z_0) \quad (5.20)$$

and its derivatives are given by

$$z_1^{(q)}(0) = \sum_{i,j=1}^s b_i \omega_{ij} Z_j^{(q)}(0). \quad (5.21)$$

We thus obtain the following result.

Theorem 5.7. *The numerical solution of (5.10) satisfies*

$$\begin{aligned} y_1^{(q)}|_{h=0} &= \sum_{t \in LDAT^2_y, \varrho(t)=q} \gamma(t) \sum_{i=1}^s b_i \Phi_i(t) F(t)(y_0, z_0), \\ z_1^{(q)}|_{h=0} &= \sum_{u \in LDAT^2_z, \varrho(u)=q} \gamma(u) \sum_{i,j=1}^s b_i \omega_{ij} \Phi_j(u) F(u)(y_0, z_0), \end{aligned}$$

where the coefficients γ and Φ_i are given in Theorem 5.6. \square

Order Conditions

A comparison of Theorem 5.7 with Theorem 5.5 gives

Theorem 5.8 (HLR89). *For the Runge-Kutta method (5.10) we have*

$$\begin{aligned} y(x_0 + h) - y_1 &= \mathcal{O}(h^{p+1}) \quad \text{iff} \\ \sum_{i=1}^s b_i \Phi_i(t) &= \frac{1}{\gamma(t)} \quad \text{for } t \in DAT^2_y, \varrho(t) \leq p, \\ z(x_0 + h) - z_1 &= \mathcal{O}(h^{q+1}) \quad \text{iff} \\ \sum_{i,j=1}^s b_i \omega_{ij} \Phi_j(u) &= \frac{1}{\gamma(u)} \quad \text{for } u \in DAT^2_z, \varrho(u) \leq q, \end{aligned}$$

where the coefficients γ and Φ_i are those of Theorem 5.6 and ω_{ij} is given by (5.19). \square

Remark 5.9. Let $P(x_0) = I - (f_z(g_y f_z)^{-1} g_y)(y_0, z_0)$ be the projection introduced in Definition 4.3. Since $P(x_0) f_z(y_0, z_0) = 0$ we have

$$P(x_0) F(t)(y_0, z_0) = 0 \quad (5.22)$$

for all trees $t \in DAT2_y$ of the form $t = [u]_y$ with $u \in DAT2_z$. Consequently, such trees of order p need not be considered for the construction of Runge-Kutta methods of order p (see Theorem 4.5).










Applying repeatedly the definition of Φ_i in Theorem 5.6 we get the following algorithm:

Forming the Order Condition for a Given Tree. Attach to each vertex one summation index; if the root is fat, attach three indices to this root. Then the left hand side of the order condition is a sum over all indices of a product with factors

- b_i if “ i ” is the index of a meagre root;
- $b_i \omega_{ij} \omega_{jk}$ if “ i, j, k ” are the three indices of a fat root;
- a_{ij} if “ j ” lies directly above “ i ” and “ j ” is meagre;
- ω_{ij} if “ j ” lies directly above “ i ” and “ j ” is fat.

In Table 5.1 we collect the order conditions for some trees of $DAT2$. We have not included the trees which have only meagre vertices, because their order condition is exactly the same as that of Sect. II.2 (Table 2.2). Several trees of $DAT2$ lead to the same order condition (Exercise 2). We also observe that some of the order conditions for the trees $[u]_y$ with $u \in DAT2_z$ are identical to those for index 1 problems (see Exercise 1 of Sect. VI.4).

Table 5.1. Trees and order conditions

$q(t)$	graph	order condition
2		$\sum b_i \omega_{ij} c_j^2 = 1$
3		$\sum b_i \omega_{ij} c_j^3 = 1$
3		$\sum b_i \omega_{ij} c_j a_{jk} c_k = \frac{1}{2}$
3		$\sum b_i c_i \omega_{ij} c_j^2 = \frac{2}{3}$
3		$\sum b_i \omega_{ij} c_j^2 \omega_{ik} c_k^2 = \frac{4}{3}$
$q(u)$	graph	order condition
1		$\sum b_i \omega_{ij} \omega_{jk} c_k^2 = 2$
2		$\sum b_i \omega_{ij} \omega_{jk} c_k^3 = 3$
2		$\sum b_i \omega_{ij} \omega_{jk} c_k a_{k\ell} c_\ell = \frac{3}{2}$
2		$\sum b_i \omega_{ij} c_j \omega_{jk} c_k^2 = 2$

Simplifying Assumptions

For the construction of implicit Runge-Kutta methods the simplifying conditions $B(p)$, $C(\eta)$, $D(\xi)$ of Sect. IV.5 play an important role. The following result extends Theorem IV.5.1 to index 2 problems.

Theorem 5.10 (HLR89, p. 67). *Suppose that the Runge-Kutta matrix (a_{ij}) is invertible and that $b_i = a_{si}$ for $i = 1, \dots, s$. Then the conditions $B(p)$, $C(\eta)$, $D(\xi)$ with $p \leq 2\eta$ and $p \leq \eta + \xi + 1$ imply that the y -component of the local error of (5.1) satisfies*

$$y_1 - y(x_0 + h) = \mathcal{O}(h^{p+1}).$$

Proof. We just outline the main ideas; details are given in (HLR89, pp. 64-67). As in Sect. II.7 (Fig. II.7.1) we first simplify the order conditions with the help of $C(\eta)$. This implies that trees with a branch ending with $[\tau, \dots, \tau]_y$ (the number of τ 's is $k-1$) where $k \leq \eta$ need no longer be considered. If we write $C(\eta)$ in the form

$$\sum_{j=1}^s \omega_{ij} c_j^k = k c_i^{k-1} \quad \text{for } k = 1, \dots, \eta, \quad (5.23)$$

we observe that trees ending with $[\tau, \dots, \tau]_z$ can also be reduced if the number of τ 's is between 1 and η .

The simplifying condition $D(\xi)$ allows us to remove trees $[\tau, \dots, \tau, t]_y$ with $t \in \text{DAT}_y$, where the number of τ 's is $\leq \xi$. Writing $D(\xi)$ as

$$\sum_{i=1}^s b_i c_i^k \omega_{ij} = \sum_{i=1}^s b_i \omega_{ij} - k b_j c_j^{k-1} \quad \text{for } k = 1, \dots, \xi \quad (5.24)$$

it follows that the trees $[\tau, \dots, \tau, u]_y$ with $u \in \text{DAT}_z$ (number of τ 's is k) can also be eliminated for $1 \leq k \leq \xi$. Since $p \leq 2\eta$ and $p \leq \eta + \xi + 1$ all that remains after these reductions are the bushy trees $[\tau, \dots, \tau]_y$ whose order conditions are satisfied by $B(p)$, and trees of the form $[u]_y$ with $u \in \text{DAT}_z$. Because of the assumption $b_i = a_{si}$ we have

$$\sum_{i=1}^s b_i \omega_{ij} = \begin{cases} 0 & \text{if } j = 1, \dots, s-1 \\ 1 & \text{if } j = s, \end{cases} \quad (5.25)$$

and these trees can also be reduced to the bushy trees. □

Remark. If the function f of (5.1a) is linear in z , i.e.,

$$f(y, z) = f_0(y) + f_z(y)z, \quad (5.26)$$

then the elementary differentials for trees $[t_1, \dots, t_m, u_1, \dots, u_n]_y$ with $n \geq 2$ vanish identically and the corresponding order conditions need not be considered.

In this situation the assumption $p \leq 2\eta$ can be relaxed to $p \leq 2\eta + 1$. An important class of problems satisfying (5.26) are constrained mechanical systems in the index 2 formulation (1.46a,b,d).

As an illustration of Theorem 5.10 we consider the Lobatto IIIC methods. They satisfy $B(p), C(\eta), D(\xi)$ with $p = 2s - 2$, $\eta = s - 1$ and $\xi = s - 1$ (see Table IV.5.13) and also $a_{si} = b_i$. It therefore follows from Theorem 5.10 that the local error satisfies $\delta y_h(x) = \mathcal{O}(h^{2s-1})$.

The following result shows that for methods which do not satisfy $a_{si} = b_i$ it is unlikely that the estimates of Lemma 4.4 can be improved.

Lemma 5.11. *Let p be the largest integer such that the y -component of the local error satisfies*

$$\delta y_h(x) = \mathcal{O}(h^{p+1}).$$

If the Runge-Kutta matrix is invertible and $c_i \neq 1$ for all i , then

$$p \leq s^*$$

where s^ is the number of distinct non-zero values among c_1, \dots, c_s .*

Proof. The order conditions for the trees $[[\tau, \dots, \tau]_z]_y$ imply that

$$\sum_{i,j=1}^s b_i \omega_{ij} \int_0^{c_j} q(t) dt = \int_0^1 q(t) dt \quad (5.27)$$

for all polynomials $q(t)$ of degree $\leq p - 1$. Put $q(t) = d'(t)$, where $d(t)$ is a polynomial of minimal degree such that $d(c_i) = 0$ for all i , $d(0) = 0$ and $d(1) \neq 0$. Condition (5.27) is violated by this polynomial. The inequality $p \leq s^*$ now follows because the degree of this polynomial $q(t)$ is s^* . \square

Projected Runge-Kutta Methods

It is, of course, interesting to study the convergence order of projected Runge-Kutta methods (Definition 4.11) which are not yet covered by Theorem 4.12. The main tool for the subsequent study is the following interpretation of projected Runge-Kutta methods.

Table 5.2. Original and extended Runge-Kutta methods

c	A	c	A	0
	b^T	$1+\epsilon$	b^T	ϵ
			b^T	ϵ

Lemma 5.12 (Lubich 1991). *Consider an s -stage Runge-Kutta method with invertible coefficient matrix A and the extended $(s+1)$ -stage method defined in Table 5.2. For an initial value y_0 satisfying $g(y_0) = 0$ denote their numerical solutions after one step by y_1 and y_1^ε , respectively. If the function f in (5.1a) is linear in z (i.e., (5.26) is satisfied), then the numerical solution \hat{y}_1 of the projected Runge-Kutta method (4.1), (4.38) satisfies*

$$\hat{y}_1 - y_1^\varepsilon = \mathcal{O}(h\varepsilon) \quad (5.28)$$

for h sufficiently small and $\varepsilon \rightarrow 0$.

Proof. The last stage of the extended $(s+1)$ -stage Runge-Kutta method reads

$$\begin{aligned} Y_{s+1} &= y_1 + h\varepsilon f(Y_{s+1}, Z_{s+1}) \\ 0 &= g(Y_{s+1}) \end{aligned} \quad (5.29)$$

and we have $y_1^\varepsilon = Y_{s+1}$ (note that this is the result of an implicit Euler step with step size $h\varepsilon$ starting from y_1). Using the linearity of f with respect to z and putting $\lambda = h\varepsilon Z_{s+1}$ we obtain

$$\begin{aligned} y_1^\varepsilon &= y_1 + h\varepsilon f_0(y_1^\varepsilon) + f_z(y_1^\varepsilon)\lambda \\ 0 &= g(y_1^\varepsilon). \end{aligned} \quad (5.30)$$

Comparing (5.30) with (4.38) the implicit function theorem implies that (5.28) is satisfied for sufficiently small h and ε . \square

The implicit function theorem, applied to (5.30), also shows that y_1^ε is as often differentiable with respect to h and ε as the right-hand side of the problem (5.1) is. Hence, the Taylor series expansion of y_1^ε with respect to h has coefficients which converge to a finite limit as $\varepsilon \rightarrow 0$.

The order conditions for a projected Runge-Kutta method (applied to (5.1), (5.26)) can thus be obtained by considering the limit $\varepsilon \rightarrow 0$ in the order conditions for the extended Runge-Kutta method (Exercise 5). Let us illustrate this by extending the statement of Theorem 5.10 to projected Runge-Kutta methods.

Theorem 5.13 (Lubich 1991). *Suppose that the Runge-Kutta matrix A is invertible and that the index 2 problem satisfies (5.26). Then the conditions $B(p)$, $C(\eta)$, $D(\xi)$ with $p \leq 2\eta + 1$ and $p \leq \eta + \xi + 1$ imply that the local error of the projected Runge-Kutta method satisfies*

$$\hat{y}_1 - y(x_0 + h) = \mathcal{O}(h^{p+1}). \quad (5.31)$$

If in addition $p \leq 2\eta$ then (5.31) holds also when f is nonlinear in z .

Proof. One verifies that the conditions $B(p)$, $C(\eta)$, $D(\xi)$, (5.23), (5.24) and (5.25) are, in the limit $\varepsilon \rightarrow 0$, also satisfied for the extended method of Table 5.2. Let us demonstrate this for the Condition (5.23). The inverse of the extended

Runge-Kutta matrix is given by

$$\begin{pmatrix} A & 0 \\ b^T & \varepsilon \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & 0 \\ -\varepsilon^{-1}b^T A^{-1} & \varepsilon^{-1} \end{pmatrix}. \quad (5.32)$$

Therefore (5.23) is seen to be satisfied for $i = 1, \dots, s$. For $i = s + 1$ one gets

$$\sum_{j=1}^{s+1} \omega_{s+1,j} c_j^k = -\varepsilon^{-1} \sum_{i,j=1}^s b_i \omega_{ij} c_j^k + \varepsilon^{-1} (1 + \varepsilon)^k. \quad (5.33)$$

Using (5.23) for $i \leq s$ and $B(p)$ the right-hand expression of (5.33) becomes $-\varepsilon^{-1} + \varepsilon^{-1} (1 + \varepsilon)^k$ and tends to k for $\varepsilon \rightarrow 0$. Hence, Condition (5.23) is, in the limit $\varepsilon \rightarrow 0$, also satisfied for $i = s + 1$. As in the proof of Theorem 5.10 (see also the remark after that proof) we deduce the statement for the case where $f(y, z)$ is linear in z .

The generalization to nonlinear problems can be proved by a perturbation argument. We let $z(x)$ be the exact solution of (5.1) and consider the problem (Lubich 1991)

$$\begin{aligned} u' &= f(u, z(x)) + f_z(u, z(x))\lambda \\ 0 &= g(u) \end{aligned} \quad (5.34)$$

in the variables u and λ . This new problem is of index 2 again and has obviously the solution $u(x) = y(x)$ and $\lambda(x) = 0$. Since (5.34) is linear in the algebraic variable λ , the theorem can be applied and we get for the projected Runge-Kutta solution

$$\hat{u}_1 - y(x_0 + h) = \mathcal{O}(h^{p+1}). \quad (5.35)$$

We still have to estimate $\hat{y}_1 - \hat{u}_1$. This is possible with the help of Theorem 4.2. In addition to the nonlinear system (4.2) (with $\eta = y_0$) we consider the method applied to (5.34):

$$\begin{aligned} U_i &= y_0 + h \sum_{j=1}^s a_{ij} \left(f(U_j, z(x_0 + c_j h)) + f_z(U_j, z(x_0 + c_j h)) \Lambda_j \right) \\ 0 &= g(U_i). \end{aligned} \quad (5.36)$$

Its first line can be written as

$$U_i = y_0 + h \sum_{j=1}^s a_{ij} f(U_j, z(x_0 + c_j h) + \Lambda_j) + \mathcal{O}(h \|\Lambda\|^2)$$

where $\|\Lambda\| = \max_j \|\Lambda_j\|$. Theorem 4.2 thus yields

$$\|U_i - Y_i\| \leq Ch \|\Lambda\|^2 \quad (5.37a)$$

$$\|\Lambda_i + z(x_0 + c_i h) - Z_i\| \leq C \|\Lambda\|^2. \quad (5.37b)$$

Since $C(\eta)$ holds, the estimate (4.14) together with (5.37b) proves $\Lambda_i = \mathcal{O}(h^\eta)$. We therefore obtain $y_1 - u_1 = \mathcal{O}(h^{2\eta+1})$ with the help of (5.37), and $\hat{y}_1 - \hat{u}_1 = \mathcal{O}(h^{2\eta+1})$ as a consequence of $z_1 - z(x_0 + h) = \mathcal{O}(h^\eta)$. \square

Examples. 1) Collocation methods satisfy $B(p)$, $C(s)$ and $D(p-s)$ where s is the number of stages and p the order of the underlying quadrature formula (consult Lemma IV.5.4). Hence, the above presentation provides an alternative proof of Theorem 4.12.

2) The projected s -stage Radau IA method (see Table IV.5.13) has order $2s-1$ for problems which are linear in z , and order $2s-2$ for general nonlinear index 2 problems.

Exercises

1. Denote by r the largest number such that the local error of the z -component satisfies $\delta z_h(x) = \mathcal{O}(h^r)$. For implicit Runge-Kutta methods with invertible coefficient matrix, $R(\infty) = 0$ and $c_j \leq 1$ (all j) prove that

$$r \leq s^*$$

where s^* is the number of distinct non-zero values among c_1, \dots, c_s .

Hint. The order conditions for the bushy trees $[\tau, \dots, \tau]_z$ imply that

$$\sum_{i,j,k} b_i \omega_{ij} \omega_{jk} \int_0^{c_k} q(t) dt = q(1)$$

for all polynomials $q(t)$ of degree $\leq r-1$.

2. If a tree of *DAT2* satisfies one of the following two conditions
 - a) a fat vertex (different from the root) is singly branched
 - b) a singly branched meagre vertex (\neq root) is followed by a fat vertex
 then the corresponding order condition is equivalent to that of a tree of the same order but with fewer fat vertices. Consequently, trees satisfying either (a) or (b) need not be considered for the construction of Runge-Kutta methods.
3. Suppose that the function $f(y, z)$ in (5.1) is linear in z . Characterize the trees of *DAT2* for which the elementary differentials vanish identically.
4. With the help of Theorem 5.10 and Lemma IV.5.4 give a new (algebraic) proof of Theorem 4.9.
5. (Lubich 1991). Consider a projected Runge-Kutta method for index 2 problems which are linear in z . Prove that $\hat{y}_1 - y(x_0 + h) = \mathcal{O}(h^4)$ iff the condition

$$\sum_{i,j=1}^s b_i (1 - c_i) \omega_{ij} c_j^2 = \frac{1}{3}$$

is satisfied in addition to the four order conditions already needed for ordinary differential equations.

VII.6 Half-Explicit Methods for Index 2 Systems

The methods of Sects. VII.3 and VII.4 do not use the semi-explicit structure of the differential-algebraic equation

$$y' = f(y, z), \quad 0 = g(y) \quad (6.1)$$

($y \in \mathbb{R}^n, z \in \mathbb{R}^m$) and can as well be applied to more general situations. Here we shall show how this structure can be exploited for the derivation of new, efficient integration methods. The main idea is to discretize the differential variables y in an explicit manner, and the algebraic variables z in an implicit manner.

The most simple method of this type is the half-explicit Euler method

$$y_1 = y_0 + hf(y_0, z_0) \quad (6.2a)$$

$$0 = g(y_1). \quad (6.2b)$$

Inserting (6.2a) into (6.2b) yields the nonlinear system $0 = g(y_0 + hf(y_0, z_0))$ for z_0 . It possesses a locally unique solution, if

$$g_y(y)f_z(y, z) \quad \text{is invertible} \quad (6.3)$$

at (y_0, z_0) . Once z_0 is computed, the value y_1 is determined explicitly by (6.2a).

This example shows some interesting features of half-explicit methods. Compared to the implicit Euler discretization, it can be implemented more efficiently, because the nonlinear system is of reduced dimension (m instead of $n + m$). Compared to the explicit Euler method in the mode “index reduction and projection” (see Sect. VII.2), it avoids an accurate computation of the derivative $g_y(y)$. The numerical approximation y_1 only depends on an initial value of the y -component, as does the exact solution of (6.1).

In this section we shall develop half-explicit Runge-Kutta methods, extrapolation methods, and multistep methods. They are in particular very efficient for constrained mechanical systems in their index 2 formulation, because nonlinear systems are completely avoided in this situation (see below).

Half-Explicit Runge-Kutta Methods

In HLR89, the following extension of (6.2) to explicit Runge-Kutta methods is proposed:

$$Y_i = y_0 + h \sum_{j=1}^{i-1} a_{ij} f(Y_j, Z_j), \quad i = 1, \dots, s \quad (6.4a)$$

$$0 = g(Y_i) \quad (6.4b)$$

$$y_1 = y_0 + h \sum_{i=1}^s b_i f(Y_i, Z_i), \quad (6.4c)$$

$$0 = g(y_1). \quad (6.4d)$$

We have $Y_1 = y_0$, and Eq. (6.4b) is automatically satisfied for $i = 1$, because the initial value is assumed to be consistent. We next insert Y_2 from (6.4a) into (6.4b) and obtain a nonlinear equation for Z_1 , which has a (locally) unique solution, if $a_{21} \neq 0$ and the usual index 2 assumption (6.3) is satisfied. We thus obtain Z_1 and Y_2 . The next step allows us to compute Z_2 and Y_3 , etc.

The local error and convergence properties of (6.4) are studied in HLR89 and Brasey & Hairer (1993). It turns out that the coefficients a_{ij}, b_i have to satisfy additional order conditions. As a consequence, 8 stages are needed for a 5th order method (Brasey 1992), compared to only 6 stages for classical Runge-Kutta methods (see Sect. II.5). Arnold (1995) and Murua (1995) have independently proposed a modification, which simplifies the order conditions and makes the approach more efficient. Their main idea is to introduce an explicit stage $Y_1 = y_0$, $Z_1 = z_0$, $Y_2 = y_0 + h a_{21} f(y_0, z_0)$, and to suppress the condition $g(Y_2) = 0$ in the second stage. We follow here the approach of Murua (1995), because it is slightly more general. For consistent initial values (y_0, z_0) we define

$$Y_1 = y_0, \quad Z_1 = z_0 \quad (6.5a)$$

$$Y_i = y_0 + h \sum_{j=1}^{i-1} a_{ij} f(Y_j, Z_j), \quad i = 2, \dots, s \quad (6.5b)$$

$$\hat{Y}_i = y_0 + h \sum_{j=1}^i \hat{a}_{ij} f(Y_j, Z_j), \quad 0 = g(\hat{Y}_i), \quad i = 2, \dots, s \quad (6.5c)$$

$$y_1 = \hat{Y}_s. \quad (6.5d)$$

The value z_1 can either be computed from the hidden constraint $g_y(y_1) f(y_1, z_1) = 0$, or from the additional stage

$$\hat{Y}_{s+1} = y_0 + h \sum_{j=1}^{s+1} \hat{a}_{s+1,j} f(Y_j, Z_j), \quad 0 = g(\hat{Y}_{s+1}) \quad (6.5e)$$

as $z_1 = Z_{s+1}$. Here we have put $Y_{s+1} = y_1$, so that the value $f(Y_{s+1}, Z_{s+1})$

can be reused as $f(y_0, z_0)$ for the next step. A significant difference compared to the original approach (6.4) is that the numerical solution (y_1, z_1) depends on both initial values $(y_0$ and $z_0)$.

Existence of the Numerical Solution. Suppose that the initial values satisfy $g(y_0) = 0$ and $g_y(y_0)f(y_0, z_0) = \mathcal{O}(\delta)$ with some sufficiently small $\delta > 0$ (we have to admit small perturbations in the hidden constraint, because in general the approximation z_1 of (6.5e) does not satisfy $g_y(y_1)f(y_1, z_1) = 0$). By an induction argument we assume that the values (Y_j, Z_j) are already known for $j = 1, \dots, i-1$, and satisfy $Y_j = y_0 + \mathcal{O}(h)$, $Z_j = z_0 + \mathcal{O}(h + \delta)$. Then, Y_i is explicitly given by (6.5b), and we have $Y_i = y_0 + \mathcal{O}(h)$. As in (3.13) we now write the condition $0 = g(\hat{Y}_i)$ as

$$0 = \int_0^1 g_y(y_0 + \tau(\hat{Y}_i - y_0)) d\tau \cdot \sum_{j=1}^i \hat{a}_{ij} f(Y_j, Z_j), \quad (6.6)$$

where \hat{Y}_i has to be replaced by (6.5c). This is a nonlinear equation of the form $F(Z_i, h) = 0$. Since $F(z_0, 0) = \mathcal{O}(\delta)$ and

$$\frac{\partial F}{\partial z}(z_0, 0) = \hat{a}_{ii} \cdot g_y(y_0) f_z(y_0, z_0),$$

it follows from the Implicit Function Theorem that (6.6) has a locally unique solution, if (6.3) and the condition

$$\hat{a}_{ii} \neq 0 \quad \text{for all } i \quad (6.7)$$

hold. Moreover we have $Z_i = z_0 + \mathcal{O}(h + \delta)$.

Error Propagation and Convergence. For inconsistent initial values we replace the nonlinear equation in (6.5c) by $g(\hat{Y}_i) = g(y_0)$, so that the method is well-defined in a whole neighbourhood of the solution manifold (observe that the above existence result is still valid). Such an extension has the advantage that differentiation with respect to initial values is possible. The method (6.5) with z_1 from (6.5e), can thus be written as

$$\begin{aligned} y_{n+1} &= y_n + h\Phi(y_n, z_n, h) \\ z_{n+1} &= \Psi(y_n, z_n, h) \end{aligned} \quad (6.8)$$

with smooth functions Φ and Ψ . For the study of convergence and, in particular, of the order conditions the triangular matrix

$$W = (w_{ij})_{i,j=1}^{s+1} = \begin{pmatrix} 1 & & & \\ \hat{a}_{21} & \hat{a}_{22} & & \\ \vdots & \vdots & \ddots & \\ \hat{a}_{s+1,1} & \hat{a}_{s+1,2} & \cdots & \hat{a}_{s+1,s+1} \end{pmatrix}^{-1} \quad (6.9)$$

will play an important role.

Lemma 6.1. *Suppose that the method (6.5), satisfying (6.7), is written in the form (6.8). If $g(y_0) = 0$ and $g_y(y_0)f(y_0, z_0) = \mathcal{O}(h)$, it holds*

$$\frac{\partial \Phi}{\partial z}(y_0, z_0, h) = \mathcal{O}(h), \quad \frac{\partial \Psi}{\partial z}(y_0, z_0, h) = w_{s+1,1} \cdot I + \mathcal{O}(h),$$

where $w_{s+1,1}$ is given by (6.9).

Proof. From (6.5b) it follows that $\partial Y_i / \partial z_0 = \mathcal{O}(h)$. Differentiation of (6.5c) with respect to z_0 thus yields

$$\frac{\partial \hat{Y}_i}{\partial z_0} = h \sum_{j=1}^i \hat{a}_{ij} f_z(y_0, z_0) \frac{\partial Z_j}{\partial z_0} + \mathcal{O}(h^2), \quad (6.10a)$$

$$0 = g_y(y_0) \frac{\partial \hat{Y}_i}{\partial z_0} + \mathcal{O}(h^2). \quad (6.10b)$$

Inserting (6.10a) into (6.10b) and multiplying with the inverse of the matrix $g_y(y_0)f_z(y_0, z_0)$, gives the relation

$$\sum_{j=1}^i \hat{a}_{ij} \frac{\partial Z_j}{\partial z_0} = \mathcal{O}(h) \quad \text{for } i = 2, \dots, s+1.$$

The statement now follows from $Z_1 = z_0$, i.e., $\partial Z_1 / \partial z_0 = I$. \square

Consider two pairs of initial values (y_0, z_0) , $(\tilde{y}_0, \tilde{z}_0)$, satisfying $g(y_0) = 0$, $g(\tilde{y}_0) = 0$, $g_y(y_0)f(y_0, z_0) = \mathcal{O}(h)$, $g_y(\tilde{y}_0)f(\tilde{y}_0, \tilde{z}_0) = \mathcal{O}(h)$. It follows from Lemma 6.1 that the differences $\Delta y_0 = y_0 - \tilde{y}_0$, \dots satisfy the recursion

$$\begin{pmatrix} \|\Delta y_1\| \\ \|\Delta z_1\| \end{pmatrix} \leq \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h^2) \\ \mathcal{O}(1) & |w_{s+1,1}| + \mathcal{O}(h) \end{pmatrix} \begin{pmatrix} \|\Delta y_0\| \\ \|\Delta z_0\| \end{pmatrix}. \quad (6.11)$$

The local error of the method (6.5) is defined as usual. We let (y_1, z_1) be the numerical approximation for initial values $(y(x), z(x))$ on the exact solution of (6.1), and denote it by $\delta y_h(x) = y_1 - y(x+h)$, $\delta z_h(x) = z_1 - z(x+h)$.

Theorem 6.2 (Murua 1995). *Consider the problem (6.1) with consistent initial values. Suppose that (6.7) holds and that*

$$|w_{s+1,1}| < 1, \quad (6.12)$$

where $w_{s+1,1}$ is given in (6.9). If the local error satisfies

$$\delta y_h(x) = \mathcal{O}(h^{r+1}), \quad \delta z_h(x) = \mathcal{O}(h^m), \quad (6.13)$$

then we have for $x_n - x_0 \leq \text{Const}$

$$y_n - y(x_n) = \mathcal{O}(h^{\min(r, m+1)}), \quad z_n - z(x_n) = \mathcal{O}(h^{\min(r, m)}).$$

Proof. The recursion (6.11) allows us to apply Lemma VI.3.9 with $\varepsilon = h^2$ and $\alpha = |w_{s+1,1}| + \mathcal{O}(h)$. This shows that the contribution of the local error at x_i to the global error at x_n is bounded by

$$C(\|\delta y_h(x_i)\| + h^2 \|\delta z_h(x_i)\|), \quad C(\|\delta y_h(x_i)\| + (h^2 + \alpha^{n-1-i}) \|\delta z_h(x_i)\|)$$

for the y - and z -component, respectively. Summing up these contributions proves the statement. \square

Order Conditions. The order conditions for method (6.5) can be derived in the same way as for Runge-Kutta methods (previous section). The only difference is that at some places the coefficients a_{ij} have to be replaced by \hat{a}_{ij} . Since $z_1 = Z_{s+1}$, the order conditions for the z -component can be directly obtained from Theorem 5.6. The result is the following:

Forming the Order Condition for a Given Tree. Attach to each vertex one summation index. Then the left-hand side of the order condition is a sum over all indices of a product with factors

\hat{a}_{si}	if “ i ” is the index of a meagre root;
$w_{s+1,i}$	if “ i ” is the index of a fat root;
a_{ij}	if the meagre vertex “ j ” lies directly above the meagre vertex “ i ”;
\hat{a}_{ij}	if the meagre vertex “ j ” lies directly above the fat vertex “ i ”;
w_{ij}	if the fat vertex “ j ” lies directly above the meagre vertex “ i ”;

The right-hand side of the order condition is the inverse of the rational number γ , defined in Theorem 5.6.

In order to satisfy the assumption (6.13) of the convergence theorem, the order conditions have to be satisfied for trees $t \in \text{DAT}^2_y$ with $\varrho(t) \leq r$, and for trees $u \in \text{DAT}^2_z$ with $\varrho(u) \leq m - 1$.

Construction of Methods. The trees of Sect. II.2 form a subset of the “index 2 trees” to be considered here. From the above construction principle it is clear that the coefficients $a_{ij}, b_i := \hat{a}_{si}$ have to satisfy the classical order conditions of Sect. II.2. It is therefore natural to take a known, explicit Runge-Kutta method of a certain order and to determine \hat{a}_{ij} in such a way that the remaining order conditions are satisfied. Arnold (1995) and Murua (1995) have shown how half-explicit methods, based on the Dormand & Prince pair of Table II.5.2, can be constructed. Let us outline the main idea.

A significant simplification of the order conditions is obtained by requiring

$$\sum_{j=1}^i \hat{a}_{ij} c_j^{q-1} = \frac{\hat{c}_i^q}{q} \quad \text{for } i = 1, \dots, s+1, \quad (6.14)$$

where $c_i = \sum_j a_{ij}$ and $\hat{c}_i = \sum_j \hat{a}_{ij}$. For $i = 1$, the relation (6.14) is automatically fulfilled because of $\hat{a}_{1j} = 0$. For $i > 1$, it can be satisfied for $q = 1$ (definition

of \hat{c}_i , $q = 2$, and $q = 3$. The simplification in the order conditions is similar to that illustrated in Fig. II.5.2. By the definition of the matrix W , the relations of Eq. (6.14) are equivalent to

$$\sum_{j=1}^i w_{ij} \hat{c}_j^q = q c_i^{q-1} \quad \text{for } i = 1, \dots, s+1. \quad (6.15)$$

This implies further reductions in the set of order conditions. The few remaining ones can be treated in a straight-forward manner. For further details and for the coefficients of the resulting method we refer to the original article of Murua (1995). They have been incorporated in the code PHEM56 (see Sect. VII.7).

Application to Constrained Mechanical Systems. Consider the system

$$q' = u \quad (6.16a)$$

$$M(q)u' = f(q, u) - G^T(q)\lambda \quad (6.16b)$$

$$0 = g(q), \quad (6.16c)$$

where $G(q) = g_q(q)$. Differentiating the constraint (6.16c) yields

$$0 = G'(q)u. \quad (6.16d)$$

If $M(q)$ is invertible, the system (6.16a,b,d) is of the form (6.1) with $y = (q, u)$ and $z = \lambda$. The assumption (6.3) is equivalent to (1.47).

For this particular system the method (6.5) can be applied as follows: assume that Q_j, U_j, Λ_j , and $U_j' = M(Q_j)^{-1}(f(Q_j, U_j) - G^T(Q_j)\Lambda_j)$ are already given for $j = 1, \dots, i-1$. We then put

$$Q_i = q_0 + h \sum_{j=1}^{i-1} a_{ij} U_j, \quad U_i = u_0 + h \sum_{j=1}^{i-1} a_{ij} U_j',$$

and compute Λ_i, U_i' from the system

$$\begin{pmatrix} M(Q_i) & G^T(Q_i) \\ G(\hat{Q}_i) & 0 \end{pmatrix} \begin{pmatrix} U_i' \\ \Lambda_i \end{pmatrix} = \begin{pmatrix} f(Q_i, U_i) \\ R_i \end{pmatrix}, \quad (6.17)$$

where $\hat{Q}_i = q_0 + h \sum_{j=1}^i \hat{a}_{ij} U_j$ and $R_i = -G(\hat{Q}_i)(u_0 + h \sum_{j=1}^{i-1} \hat{a}_{ij} U_j') / (h \hat{a}_{ii})$ are known quantities. Hence, only linear systems of type (6.17) have to be solved. This makes half-explicit methods very attractive for the numerical solution of constrained mechanical systems. If necessary, this method can be combined with projections as explained in Sect. VII.2, so that also the position constraint is satisfied by the numerical approximation.

We remark that the methods proposed by Arnold (1995) satisfy $\hat{Q}_i = Q_{i+1}$ for $i \geq 2$, so that some G evaluations can be saved in the computation of (6.17).

Extrapolation Methods

For nonstiff ordinary differential equations, the most efficient extrapolation algorithm is the GBS method (see Sect. II.9). Lubich (1989) extends this method to differential-algebraic equations of index 2.

Consider an initial value y_0 satisfying $g(y_0) = 0$. Then, an approximation $S_h(x)$ to $y(x)$ (with $x = x_0 + 2mh$) is defined by

$$y_1 = y_0 + hf(y_0, z_0), \quad g(y_1) = 0 \quad (6.18a)$$

$$y_{i+1} = y_{i-1} + 2hf(y_i, z_i), \quad g(y_{i+1}) = 0, \quad i = 1, \dots, 2m \quad (6.18b)$$

$$S_h(x) = (y_{2m-1} + 2y_{2m} + y_{2m+1})/4. \quad (6.18c)$$

The starting step is identical to the half-explicit Euler method, considered at the beginning of this section. It is implicit in z_0 and explicit in y_1 . For the case that Eq. (6.1) is linear in z , i.e.,

$$f(y, z) = f_0(y) + f_z(y)z, \quad (6.19)$$

we shall show below that the numerical approximations $S_h(x_0 + 2mh)$ and z_{2m} possess an h^2 -expansion. Hence, these values can be used as the basis of an extrapolation method. The implementation is completely analogue to that for the GBS method (choice of the step number sequence, order and step size control, dense output, ...). Since the extrapolated values do not satisfy the constraint $g(y) = 0$, it is recommended to project them onto this manifold (as explained in Sect. VII.2) after every accepted step.

The assumption (6.19) is satisfied for many interesting problems, e.g., for the constrained mechanical system (6.16a,b,d), where $z = \lambda$ plays the role of a Lagrange multiplier.

Theorem 6.3 (Lubich 1989). *Under the assumptions (6.3) and (6.19) the numerical solution of method (6.18) possesses an asymptotic h^2 -expansion*

$$y_{2m} - y(x_{2m}) = a_2(x_{2m})h^2 + a_4(x_{2m})h^4 + \dots + a_{2N}(x_{2m})h^{2N} + \mathcal{O}(h^{2N+2})$$

$$z_{2m} - z(x_{2m}) = b_2(x_{2m})h^2 + b_4(x_{2m})h^4 + \dots + b_{2N}(x_{2m})h^{2N} + \mathcal{O}(h^{2N+2})$$

and another h^2 -expansion for the error of $S_h(x_{2m})$.

The numerical solution $\{y_i\}$ of method (6.18) lies on the manifold defined by $g(y) = 0$. In order to be able to apply the results and ideas of Sects. II.8 and II.9, we extend the method (6.18) to arbitrary initial values as follows:

$$y_1 = y_0 + hf(y_0, z_0), \quad g(y_1) = g(y_0) \quad (6.20a)$$

$$y_{i+1} = y_{i-1} + 2hf(y_i, z_i), \quad g(y_{i+1}) = g(y_{i-1}), \quad i = 1, \dots, 2m \quad (6.20b)$$

We further eliminate the z -variables: using the identity

$$g(y_{i+1}) - g(y_{i-1}) = \int_{-1}^1 g_y \left(\frac{y_{i+1} + y_{i-1}}{2} + \sigma \frac{y_{i+1} - y_{i-1}}{2} \right) d\sigma \cdot \left(\frac{y_{i+1} - y_{i-1}}{2} \right),$$

Eq. (6.20b) becomes

$$0 = \int_{-1}^1 g_y \left(\frac{y_{i+1} + y_{i-1}}{2} + \sigma h f(y_i, z_i) \right) d\sigma \cdot f(y_i, z_i). \quad (6.21)$$

By assumption (6.3) and the Implicit Function Theorem, Eq. (6.21) can be solved for z_i as a smooth function of $(y_{i+1} + y_{i-1})/2$, y_i , and h . Inserted into (6.20b) we obtain a recursion of the type

$$y_{i+1} = y_{i-1} + 2h\Phi(y_i, (y_{i+1} + y_{i-1})/2, h). \quad (6.22)$$

The starting step (6.20a) can be rewritten in a similar way. We consider the more general system

$$w = v + hf(u, z), \quad g(w) = g(v), \quad (6.23)$$

where u, v , and h are given. It can be written in the equivalent form

$$0 = \int_0^1 g_y(v + \tau hf(u, z)) d\tau \cdot f(u, z),$$

which yields z as a smooth function of u, v , and h (again by the Implicit Function Theorem). Hence, the solution of (6.23) can be written as

$$w = v + h\Phi_0(u, v, h), \quad (6.24)$$

and the starting step (6.20a) becomes

$$y_1 = y_0 + h\Phi_0(y_0, y_0, h). \quad (6.25)$$

The crucial point of these reformulations is that the two-step method (6.22) and the starting step (6.25) are not only defined on the manifold $g(y) = 0$, but on an open neighbourhood of it. Therefore, the standard ODE theory can be applied. Results for the method (6.22), (6.25) immediately carry over to the method (6.18), because both methods are identical for initial values satisfying $g(y_0) = 0$.

Asymptotic Expansion for Symmetric Two-Step Methods. Motivated by the above reformulations we consider the method

$$y_1 = y_0 + h\Phi_0(y_0, y_0, h) \quad (6.26a)$$

$$y_{i+1} = y_{i-1} + 2h\Phi(y_i, (y_{i+1} + y_{i-1})/2, h), \quad (6.26b)$$

where Φ_0 and Φ are arbitrary, smooth increment functions. We assume that $\Phi_0(y, y, 0) = \Phi(y, y, 0) = f(y)$, so that both methods are consistent with the ordinary differential equation $y' = f(y)$. In order to get an h^2 -expansion of the error, the starting step (6.26a) has to be compatible with (6.26b) in the following sense: for arbitrary u_k, v_k , the three values

$$\begin{aligned} y_{2k-1} &:= v_k - h\Phi_0(u_k, v_k, -h), & y_{2k} &:= u_k, \\ y_{2k+1} &:= v_k + h\Phi_0(u_k, v_k, h) \end{aligned} \quad (6.27)$$

satisfy the recursion (6.26b).

Theorem 6.4. *If the method (6.26) satisfies the compatibility condition (6.27), the numerical approximations*

$$y_{2m}, \quad (y_{2m+1} + y_{2m-1})/2$$

have an asymptotic expansion in even powers of h .

Proof. Inspired by Stetter's proof of Theorem II.9.2 we put $u_k := y_{2k}$, and let v_k be the solution of

$$y_{2k+1} := v_k + h\Phi_0(u_k, v_k, h). \quad (6.28)$$

We thus get the one-step method in doubled dimension

$$\begin{pmatrix} u_{k+1} \\ v_{k+1} \end{pmatrix} = \begin{pmatrix} u_k \\ v_k \end{pmatrix} + h^* \begin{pmatrix} \Phi(y_{2k+1}, (u_{k+1} + u_k)/2, h^*/2) \\ \frac{1}{2}(\Phi_0(u_k, v_k, h^*/2) + \Phi_0(u_{k+1}, v_{k+1}, -h^*/2)) \end{pmatrix},$$

where $h^* = 2h$, and y_{2k+1} is given by (6.28). The assumption (6.27) implies that this one-step method is symmetric. Therefore, $y_{2m} = u_m$ and v_m have an asymptotic h^2 -expansion (see Theorem II.8.10). From

$$(y_{2m+1} + y_{2m-1})/2 = y_{2m} + h(\Phi_0(u_m, v_m, h) - \Phi_0(u_m, v_m, -h))$$

it follows that the same is true for $(y_{2m+1} + y_{2m-1})/2$. \square

Proof of Theorem 6.3. We have already seen that the method (6.20) can be written in the form (6.26). All that remains to do is to check the compatibility condition (6.27). By definition of $\Phi_0(u, v, h)$ (see the equivalence of Eqs. (6.23) and (6.25)) we have

$$\begin{aligned} y_{2k-1} &= v_k - hf(u_k, z^-), & g(y_{2k-1}) &= g(v_k) \\ y_{2k+1} &= v_k + hf(u_k, z^+), & g(y_{2k+1}) &= g(v_k). \end{aligned}$$

Since f is linear in z , this implies (6.20b) with $z_{2k} = (z^- + z^+)/2$. The asymptotic h^2 -expansion of y_{2m} and $S_h(x_{2m})$ thus follows from Theorem 6.4. From (6.21) we then see that also z_{2m} has an h^2 -expansion. \square

β -Blocked Multistep Methods

The convergence analysis of Sect. VII.3 shows that all roots of the σ -polynomial of a multistep method must lie inside the unit disc in order to get a convergent method of order p . This is a severe restriction and excludes, for example, all explicit and implicit Adams methods. Arévalo, Führer & Söderlind (1995) suggest a modification which allows the use of “nonstiff” multistep methods. The idea is to treat different parts of the problem by different discretizations.

For the index 2 problem

$$y' = f_0(y) + f_z(y)z, \quad 0 = g(y), \quad (6.29)$$

where $f(y, z) = f_0(y) + f_z(y)z$ depends linearly on z , we consider the discretization

$$\sum_{i=0}^k \alpha_i y_{n+i} = h \sum_{i=0}^k \beta_i f(y_{n+i}, z_{n+i}) - h f_z(y_{n+k}) \sum_{i=0}^k \gamma_i z_{n+i}, \quad (6.30)$$

with $g(y_{n+k}) = 0$, and denote the generating polynomials by

$$\varrho(\zeta) = \sum_{i=0}^k \alpha_i \zeta^i, \quad \sigma(\zeta) = \sum_{i=0}^k \beta_i \zeta^i, \quad \tau(\zeta) = \sum_{i=0}^k \gamma_i \zeta^i.$$

Theorem 6.5 (Arévalo, Führer & Söderlind 1996). *Let the index 2 problem (6.29) satisfy (6.3). Assume that the multistep method (ϱ, σ) is stable and of order p ($p = k$ or $p = k + 1$), that $\tau(\zeta) = \gamma_k(\zeta - 1)^k$, and that all roots of $\sigma(\zeta) - \tau(\zeta)$ lie inside the unit disc $|\zeta| < 1$. Then the global error satisfies for $x_n - x_0 \leq \text{Const}$*

$$y_n - y(x_n) = \mathcal{O}(h^p), \quad z_n - z(x_n) = \mathcal{O}(h^k).$$

Proof. The special form of $\tau(\zeta)$ is equivalent to

$$\sum_{i=0}^k \gamma_i z(x_n + ih) = \mathcal{O}(h^k),$$

so that the newly added term in (6.30) is small. Moreover, this term is premultiplied by $f_z(y_{n+k})$, so that the local error satisfies

$$\delta y_h(x) = \mathcal{O}(h^{k+1}), \quad P(x) \delta y_h(x) = \mathcal{O}(h^{p+1}),$$

where $P(x)$ is the projector of Definition 4.3.

With these observations in mind, the convergence result is obtained along the lines of the proof of Theorem 3.6. The only difference is that the coefficients β_i have to be replaced by $\beta_i - \gamma_i$ in Eqs. (3.43) and (3.44). \square

In principle, one can take any convergent multistep method (ϱ, σ) of order $p = k$ or $p = k + 1$, and try to optimize the parameter γ_k in $\tau(\zeta)$ in such a way that the roots of $\sigma(\zeta) - \tau(\zeta)$ become small. The result, for the implicit Adams methods, is rather disappointing. Only for $k \leq 3$ it is possible to obtain convergent β -blocked Adams methods (Arévalo, Führer & Söderlind (1996), see also Exercise 3).

Difference Corrected BDF. Consider the $(k + 1)$ -step BDF method, defined in Eq. (III.1.22'), and replace $\nabla^{k+1} y_{n+1}$ by $\nabla^k f_{n+1}$. This leads to the so-called difference corrected BDF

$$\sum_{j=1}^k \frac{1}{j} \nabla^j y_{n+1} = h \left(f_{n+1} - \frac{1}{k+1} \nabla^k f_{n+1} \right), \quad (6.31)$$

introduced by Söderlind (1989). Method (6.31) is a k -step method of order $p = k + 1$. Its ϱ -polynomial is identical to that of the BDF method and $\sigma(\zeta) = \zeta^k - (\zeta - 1)^k / (k + 1)$. With $\tau(\zeta) = -(\zeta - 1)^k / (k + 1)$ the difference $\sigma(\zeta) - \tau(\zeta)$ has all roots equal to zero. This is therefore an ideal candidate for a method of type (6.30).

Exercises

1. Construct all half-explicit methods (6.5) of order 3 ($r = m = 3$ in Eq. (6.13)) with $s = 3$ stages. You can take $c_2, c_3, \alpha, \widehat{c}_2, \widehat{c}_4$ as free parameters.

Hint. Start with a classical Runge-Kutta method of order 3 (Exercise 4 of Sect. II.1), and show that the order conditions imply (6.14) for $q = 2$.

2. Show that the method (IV.9.15) of Bader & Deuffhard (1983) is of the form (6.26) with

$$\begin{aligned}\Phi(u, v, h) &= f(u) - Ju + Jv \\ \Phi_0(u, v, h) &= (I - hJ)^{-1}(f(u) - Ju + Jv).\end{aligned}$$

Check the assumption (6.27).

3. Let (ϱ_k, σ_k) be the generating polynomials of the k -step implicit Adams methods (Sect. III.1). For $k = 1, 2, \dots, 10$ study numerically the function

$$R_k(\gamma) := \max \{ |\zeta^*| ; \zeta^* \text{ is root of } \sigma_k(\zeta) - \gamma(\zeta - 1)^k = 0 \}.$$

For which values of k is it possible to find γ with $R_k(\gamma) < 1$?

VII.7 Computation of Multibody Mechanisms

Dynamics of multibody systems is of great importance in the fields of robotics, biomechanics, spacecraft control, road and rail vehicle design, and dynamics of machinery.

(W. Schiehlen 1990)

After having seen several different approaches for the numerical solution of constrained mechanical systems, we are interested in their efficiency when applied to a concrete situation. We consider two particular multibody mechanisms with constraints, one nonstiff and one stiff. General references for the computation of mechanical systems are Haug (1989) and Roberson & Schwertassek (1988).

Description of the Model

We first consider “Andrews’ squeezer mechanism”, which has become prominent through the work of Giles (1978) and Manning (1981), who promoted it as a test example for numerical codes; see also Ormrod & Andrews (1986). It consists of 7 rigid bodies connected by joints without friction in plane motion. It is represented in Fig. 7.1, which we have copied (with permission) from the book of Schiehlen (1990). The numerical constants, also taken from Schiehlen (1990), are displayed in Tables 7.1 and 7.2. The arrows in the right picture of Fig. 7.1 indicate the positions of the centres of gravity C_1, \dots, C_7 . In Table 7.1 the spring coefficient of the spring connecting the point D with C is denoted by c_0 and the unstretched length is ℓ_0 . We suppose that the mechanism is driven by a motor, located at O , whose constant drive torque is given by $mom = 0.033$. The coordinate origin is the point O in Fig. 7.1 and the coordinates of the other fixed points A, B and C are given by

$$\begin{pmatrix} xa \\ ya \end{pmatrix} = \begin{pmatrix} -0.06934 \\ -0.00227 \end{pmatrix}, \begin{pmatrix} xb \\ yb \end{pmatrix} = \begin{pmatrix} -0.03635 \\ 0.03273 \end{pmatrix}, \begin{pmatrix} xc \\ yc \end{pmatrix} = \begin{pmatrix} 0.014 \\ 0.072 \end{pmatrix}. \quad (7.1)$$

Table 7.1. Geometrical parameters

$d = 0.028$	$da = 0.0115$	$e = 0.02$
$ea = 0.01421$	$zf = 0.02$	$fa = 0.01421$
$rr = 0.007$	$ra = 0.00092$	$ss = 0.035$
$sa = 0.01874$	$sb = 0.01043$	$sc = 0.018$
$sd = 0.02$	$zt = 0.04$	$ta = 0.02308$
$tb = 0.00916$	$u = 0.04$	$ua = 0.01228$
$ub = 0.00449$	$c_0 = 4530$	$\ell_0 = 0.07785$

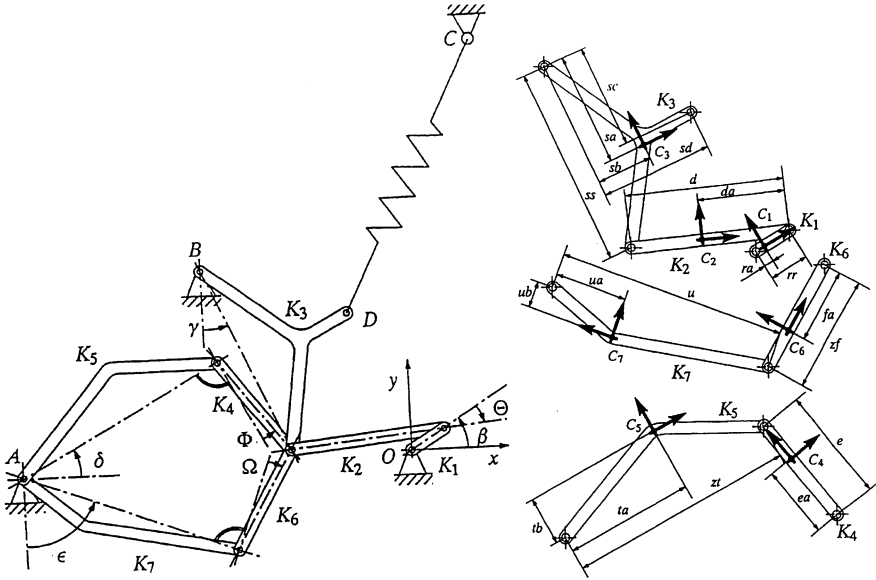


Fig. 7.1. Seven body mechanism (Schiehlen 1990, with permission)

Table 7.2. Parameters of the 7 bodies

No.	masses m_1 to m_7	inertias I_1 to I_7
1	0.04325	$2.194 \cdot 10^{-6}$
2	0.00365	$4.410 \cdot 10^{-7}$
3	0.02373	$5.255 \cdot 10^{-6}$
4	0.00706	$5.667 \cdot 10^{-7}$
5	0.07050	$1.169 \cdot 10^{-5}$
6	0.00706	$5.667 \cdot 10^{-7}$
7	0.05498	$1.912 \cdot 10^{-5}$

In order to derive the equations of motion we use the angles (see Fig. 7.1)

$$q_1 = \beta, \quad q_2 = \Theta, \quad q_3 = \gamma, \quad q_4 = \Phi, \quad q_5 = \delta, \quad q_6 = \Omega, \quad q_7 = \epsilon, \quad (7.2)$$

as position coordinates for the mechanical system. If (x_j, y_j) are the cartesian coordinates of the centre of gravity C_j ($j = 1, \dots, 7$), the *kinetic energy* of the multibody system is

$$T = \sum_{j=1}^7 m_j \frac{\dot{x}_j^2 + \dot{y}_j^2}{2} + \sum_{j=1}^7 I_j \frac{\dot{\omega}_j^2}{2} \quad (7.3)$$

where ω_j is the total angle of rotation of the j th body and m_j , I_j are constants given in Table 7.2. The values of $x_j, y_j, \dot{x}_j^2 + \dot{y}_j^2$ and $\dot{\omega}_j$ can be obtained in terms

of (7.2) by simple geometry (see Fig. 7.1):

$$C_1 : \quad x_1 = ra \cdot \cos \beta$$

$$y_1 = ra \cdot \sin \beta$$

$$\dot{x}_1^2 + \dot{y}_1^2 = ra^2 \cdot \dot{\beta}^2$$

$$\dot{\omega}_1 = \dot{\beta}$$

$$C_2 : \quad x_2 = rr \cdot \cos \beta - da \cdot \cos(\beta + \Theta)$$

$$y_2 = rr \cdot \sin \beta - da \cdot \sin(\beta + \Theta)$$

$$\begin{aligned} \dot{x}_2^2 + \dot{y}_2^2 = & (rr^2 - 2 \cdot da \cdot rr \cdot \cos \Theta + da^2) \cdot \dot{\beta}^2 \\ & + 2 \cdot (-rr \cdot da \cdot \cos \Theta + da^2) \cdot \dot{\beta} \cdot \dot{\Theta} + da^2 \cdot \dot{\Theta}^2 \end{aligned}$$

$$\dot{\omega}_2 = \dot{\beta} + \dot{\Theta}$$

$$C_3 : \quad x_3 = xb + sa \cdot \sin \gamma + sb \cdot \cos \gamma$$

$$y_3 = yb - sa \cdot \cos \gamma + sb \cdot \sin \gamma$$

$$\dot{x}_3^2 + \dot{y}_3^2 = (sa^2 + sb^2) \cdot \dot{\gamma}^2$$

$$\dot{\omega}_3 = \dot{\gamma}$$

$$C_4 : \quad x_4 = xa + zt \cdot \cos \delta + (e - ea) \cdot \sin(\Phi + \delta)$$

$$y_4 = ya + zt \cdot \sin \delta - (e - ea) \cdot \cos(\Phi + \delta)$$

$$\begin{aligned} \dot{x}_4^2 + \dot{y}_4^2 = & (e - ea)^2 \cdot \dot{\Phi}^2 + 2 \cdot ((e - ea)^2 + zt \cdot (e - ea) \cdot \sin \Phi) \cdot \dot{\Phi} \cdot \dot{\delta} \\ & + (zt^2 + 2 \cdot zt \cdot (e - ea) \cdot \sin \Phi + (e - ea)^2) \cdot \dot{\delta}^2 \end{aligned}$$

$$\dot{\omega}_4 = \dot{\Phi} + \dot{\delta}$$

$$C_5 : \quad x_5 = xa + ta \cdot \cos \delta - tb \cdot \sin \delta$$

$$y_5 = ya + ta \cdot \sin \delta + tb \cdot \cos \delta$$

$$\dot{x}_5^2 + \dot{y}_5^2 = (ta^2 + tb^2) \cdot \dot{\delta}^2$$

$$\dot{\omega}_5 = \dot{\delta}$$

$$C_6 : \quad x_6 = xa + u \cdot \sin \varepsilon + (zf - fa) \cdot \cos(\Omega + \varepsilon)$$

$$y_6 = ya - u \cdot \cos \varepsilon + (zf - fa) \cdot \sin(\Omega + \varepsilon)$$

$$\begin{aligned} \dot{x}_6^2 + \dot{y}_6^2 = & (zf - fa)^2 \cdot \dot{\Omega}^2 + 2 \cdot ((zf - fa)^2 - u \cdot (zf - fa) \cdot \sin \Omega) \cdot \dot{\Omega} \cdot \dot{\varepsilon} \\ & + ((zf - fa)^2 - 2 \cdot u \cdot (zf - fa) \cdot \sin \Omega + u^2) \cdot \dot{\varepsilon}^2 \end{aligned}$$

$$\dot{\omega}_6 = \dot{\Omega} + \dot{\varepsilon}$$

$$C_7 : \quad x_7 = xa + ua \cdot \sin \varepsilon - ub \cdot \cos \varepsilon$$

$$y_7 = ya - ua \cdot \cos \varepsilon - ub \cdot \sin \varepsilon$$

$$\dot{x}_7^2 + \dot{y}_7^2 = (ua^2 + ub^2) \cdot \dot{\varepsilon}^2$$

$$\dot{\omega}_7 = \dot{\varepsilon}$$

The *potential energy* of the system is due to the motor at the origin and to the spring connecting the point D with C . By Hooke's law it is

$$U = -mom \cdot \beta + c_0 \frac{(\ell - \ell_0)^2}{2}, \quad (7.4)$$

where ℓ is the distance between D and C , namely

$$\begin{aligned} \ell &= \sqrt{(xd - xc)^2 + (yd - yc)^2} \\ xd &= xb + sc \cdot \sin \gamma + sd \cdot \cos \gamma \\ yd &= yb - sc \cdot \cos \gamma + sd \cdot \sin \gamma. \end{aligned}$$

Finally, we have to formulate the *algebraic constraints*. The mechanism contains three loops. The first loop connects O with B via K_1, K_2, K_3 ; the other two loops connect O with A , one via K_1, K_2, K_4, K_5 , the other via K_1, K_2, K_6, K_7 . For each loop we get two algebraic conditions:

$$\begin{aligned} rr \cdot \cos \beta - d \cdot \cos(\beta + \Theta) - ss \cdot \sin \gamma &= xb \\ rr \cdot \sin \beta - d \cdot \sin(\beta + \Theta) + ss \cdot \cos \gamma &= yb \\ rr \cdot \cos \beta - d \cdot \cos(\beta + \Theta) - e \cdot \sin(\Phi + \delta) - zt \cdot \cos \delta &= xa \\ rr \cdot \sin \beta - d \cdot \sin(\beta + \Theta) + e \cdot \cos(\Phi + \delta) - zt \cdot \sin \delta &= ya \\ rr \cdot \cos \beta - d \cdot \cos(\beta + \Theta) - zf \cdot \cos(\Omega + \varepsilon) - u \cdot \sin \varepsilon &= xa \\ rr \cdot \sin \beta - d \cdot \sin(\beta + \Theta) - zf \cdot \sin(\Omega + \varepsilon) + u \cdot \cos \varepsilon &= ya. \end{aligned} \quad (7.5)$$

With the position coordinates q from (7.2) the equations (7.5) represent the constraint $g(q) = 0$ where $g: \mathbb{R}^7 \rightarrow \mathbb{R}^6$. Together with the kinetic energy T of (7.3) the potential energy U of (7.4) and $L = T - U - \lambda_1 g_1 - \dots - \lambda_6 g_6$ the equations of motion (1.46) are fully determined.

Fortran Subroutines

For the reader's convenience we include the essential parts of the FORTRAN subroutines describing the differential-algebraic problem. The equations of motion are of the form

$$M(q)\ddot{q} = f(q, \dot{q}) - G^T(q)\lambda \quad (7.6a)$$

$$0 = g(q) \quad (7.6b)$$

where $q \in \mathbb{R}^7$ is the vector defined in (7.2) and $\lambda \in \mathbb{R}^6$. In the following description the variables $Q(1), \dots, Q(7)$ correspond to $\beta, \dots, \varepsilon$ (exactly as in (7.2)) and $QP(1), \dots, QP(7)$ to their derivatives $\dot{\beta}, \dots, \dot{\varepsilon}$. In all subroutines we have used the abbreviations

SIBE = SIN (Q(1))	COBE = COS (Q(1))
SITH = SIN (Q(2))	COTH = COS (Q(2))
SIGA = SIN (Q(3))	COGA = COS (Q(3))

SIPH = SIN (Q(4))	COPH = COS (Q(4))
SIDE = SIN (Q(5))	CODE = COS (Q(5))
SIOM = SIN (Q(6))	COOM = COS (Q(6))
SIEP = SIN (Q(7))	COEP = COS (Q(7))
SIBETH = SIN (Q(1)+Q(2))	COBETH = COS (Q(1)+Q(2))
SIPHDE = SIN (Q(4)+Q(5))	COPHDE = COS (Q(4)+Q(5))
SIOMEPE = SIN (Q(6)+Q(7))	COOMEPE = COS (Q(6)+Q(7))
BEP = QP(1)	THP = QP(2)
PHP = QP(4)	DEP = QP(5)
OMP = QP(6)	EPP = QP(7)

The remaining parameters $XA, YA, \dots, D, DA, E, EA, \dots, M1, I1, M2, \dots$ are those of (7.1) and Tables 7.1 and 7.2. They usually reside in a COMMON block. The elements of $M(q)$ in (7.6) are given by

$$m_{ij} = \frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j} = \frac{\partial^2 T}{\partial \dot{q}_i \partial \dot{q}_j}.$$

This matrix is symmetric and (due to the special arrangement of the coordinates) tridiagonal. The non-zero elements (on and below the diagonal) are

```

M(1,1) = M1*RA**2 + M2*(RR**2-2*DA*RR*COTH+DA**2) + I1 + I2
M(2,1) = M2*(DA**2-DA*RR*COTH) + I2
M(2,2) = M2*DA**2 + I2
M(3,3) = M3*(SA**2+SB**2) + I3
M(4,4) = M4*(E-EA)**2 + I4
M(5,4) = M4*((E-EA)**2+ZT*(E-EA)*SIPH) + I4
M(5,5) = M4*(ZT**2+2*ZT*(E-EA)*SIPH+(E-EA)**2) + M5*(TA**2+TB**2)
+      + I4 + I5
M(6,6) = M6*(ZF-FA)**2 + I6
M(7,6) = M6*((ZF-FA)**2-U*(ZF-FA)*SIOM) + I6
M(7,7) = M6*((ZF-FA)**2-2*U*(ZF-FA)*SIOM+U**2) + M7*(UA**2+UB**2)
+      + I6 + I7

```

The i th component of the function f in (7.6) is defined by

$$f_i(q, \dot{q}) = \frac{\partial(T-U)}{\partial q_i} - \sum_{j=1}^7 \frac{\partial^2(T-U)}{\partial \dot{q}_i \partial \dot{q}_j} \cdot \dot{q}_j.$$

Written as FORTRAN statements we have

```

XD = SD*COGA + SC*SIGA + XB
YD = SD*SIGA - SC*COGA + YB
LANG = SQRT ((XD-XC)**2 + (YD-YC)**2)
FORCE = - C0 * (LANG - L0)/LANG
FX = FORCE * (XD-XC)
FY = FORCE * (YD-YC)
F(1) = MOM - M2*DA*RR*THP*(THP+2*BEP)*SITH
F(2) = M2*DA*RR*BEP**2*SITH
F(3) = FX*(SC*COGA - SD*SIGA) + FY*(SD*COGA + SC*SIGA)
F(4) = M4*ZT*(E-EA)*DEP**2*COPH
F(5) = - M4*ZT*(E-EA)*PHP*(PHP+2*DEP)*COPH
F(6) = - M6*U*(ZF-FA)*EPP**2*COOM
F(7) = M6*U*(ZF-FA)*OMP*(OMP+2*EPP)*COOM

```

The algebraic constraints $g(q) = 0$ are given by the following six equations (see (7.5))

```

G(1) = RR*COBE - D*COBETH - SS*SIGA - XB
G(2) = RR*SIBE - D*SIBETH + SS*COGA - YB
G(3) = RR*COBE - D*COBETH - E*SIPHDE - ZT*CODE - XA
G(4) = RR*SIBE - D*SIBETH + E*COPHDE - ZT*SIDE - YA
G(5) = RR*COBE - D*COBETH - ZF*COOMEF - U*SIEP - XA
G(6) = RR*SIBE - D*SIBETH - ZF*SIOMEF + U*COEP - YA

```

And here is the Jacobian matrix $G(q) = g_q(q)$. The non-zero entries of this 6×7 array are

```

GQ(1,1) = - RR*SIBE + D*SIBETH      GQ(4,2) = - D*COBETH
GQ(1,2) = D*SIBETH                  GQ(4,4) = - E*SIPHDE
GQ(1,3) = - SS*COGA                 GQ(4,5) = - E*SIPHDE - ZT*CODE
GQ(2,1) = RR*COBE - D*COBETH        GQ(5,1) = - RR*SIBE + D*SIBETH
GQ(2,2) = - D*COBETH                GQ(5,2) = D*SIBETH
GQ(2,3) = - SS*SIGA                 GQ(5,6) = ZF*SIOMEF
GQ(3,1) = - RR*SIBE + D*SIBETH       GQ(5,7) = ZF*SIOMEF - U*COEP
GQ(3,2) = D*SIBETH                  GQ(6,1) = RR*COBE - D*COBETH
GQ(3,4) = - E*COPHDE                GQ(6,2) = - D*COBETH
GQ(3,5) = - E*COPHDE + ZT*SIDE       GQ(6,6) = - ZF*COOMEF
GQ(4,1) = RR*COBE - D*COBETH        GQ(6,7) = - ZF*COOMEF - U*SIEP

```

If we apply a numerical method to the index 1 formulation of the system, we also need the expression $g_{qq}(q)(\dot{q}, \dot{q})$. It is given by

```

GQQ(1) = - RR*COBE*V(1)**2 + D*COBETH*(V(1)+V(2))**2 +
+ SS*SIGA*V(3)**2
GQQ(2) = - RR*SIBE*V(1)**2 + D*SIBETH*(V(1)+V(2))**2 -
+ SS*COGA*V(3)**2
GQQ(3) = - RR*COBE*V(1)**2 + D*COBETH*(V(1)+V(2))**2 +
+ E*SIPHDE*(V(4)+V(5))**2 + ZT*CODE*V(5)**2
GQQ(4) = - RR*SIBE*V(1)**2 + D*SIBETH*(V(1)+V(2))**2 -
+ E*COPHDE*(V(4)+V(5))**2 + ZT*SIDE*V(5)**2
GQQ(5) = - RR*COBE*V(1)**2 + D*COBETH*(V(1)+V(2))**2 +
+ ZF*COOMEF*(V(6)+V(7))**2 + U*SIEP*V(7)**2
GQQ(6) = - RR*SIBE*V(1)**2 + D*SIBETH*(V(1)+V(2))**2 +
+ ZF*SIOMEF*(V(6)+V(7))**2 - U*COEP*V(7)**2

```

Computation of Consistent Initial Values

We first compute a solution of $g(q) = 0$. Since g consists of 6 equations in 7 unknowns we can fix one of them arbitrarily, say $\Theta(0) = 0$, and compute the remaining coordinates by Newton iterations. This gives

$$\begin{aligned}
 \beta(0) &= -0.0617138900142764496358948458001 \\
 \gamma(0) &= 0.455279819163070380255912382449 \\
 \Phi(0) &= 0.222668390165885884674473185609 \\
 \delta(0) &= 0.487364979543842550225598953530 \\
 \Omega(0) &= -0.222668390165885884674473185609 \\
 \varepsilon(0) &= 1.23054744454982119249735015568.
 \end{aligned} \tag{7.7}$$

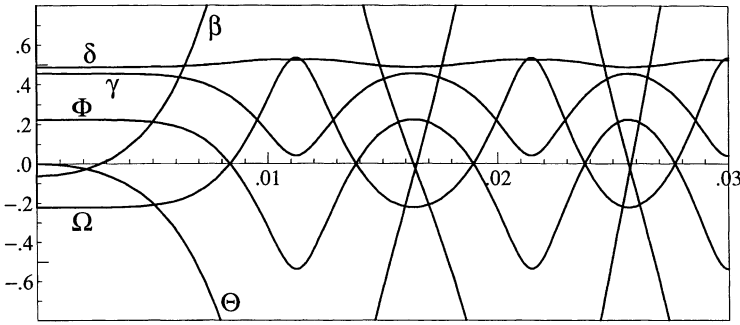


Fig. 7.2. Solution of 7 body mechanism

The condition $G(q)\dot{q} = 0$ is satisfied if we put

$$\dot{\beta}(0) = \dot{\Theta}(0) = \dot{\gamma}(0) = \dot{\Phi}(0) = \dot{\delta}(0) = \dot{\Omega}(0) = \dot{\varepsilon}(0) = 0. \quad (7.8)$$

The values of $\lambda(0)$ and $\ddot{q}(0)$ are then uniquely determined by (7.6a) and the twice differentiated constraint $0 = g_{qq}(q)(\dot{q}, \dot{q}) + G(q)\ddot{q}$. We just have to solve a linear system with the matrix

$$\begin{pmatrix} M(q) & G^T(q) \\ G(q) & 0 \end{pmatrix}. \quad (7.9)$$

Observe that g_{qq} need not be evaluated, because $\dot{q}(0) = 0$. Due to the choice $\Theta(0) = 0$ most components of $\lambda(0)$ and $\ddot{q}(0)$ vanish. Only the first two of these are different from zero and given by

$$\begin{aligned} \ddot{\beta}(0) &= 14222.4439199541138705911625887 \\ \ddot{\Theta}(0) &= -10666.8329399655854029433719415 \\ \lambda_1(0) &= 98.5668703962410896057654982170 \\ \lambda_2(0) &= -6.12268834425566265503114393122. \end{aligned} \quad (7.10)$$

The solution of this seven body mechanism is plotted (mod 2π) in Fig. 7.2 for $0 \leq t \leq 0.03$.

Numerical Computations

We first transform (7.6) into a first order system by introducing the new variable $v = \dot{q}$. Our codes apply only to problems where the derivative is multiplied by a constant matrix. We therefore also consider $w = \ddot{q}$ as a variable so that (7.6a) becomes an algebraic relation. The various formulations of the problem, as discussed in Sect. VII.1, are now as follows:

Index 3 Formulation. With $v = \dot{q}$ and $w = \ddot{q}$ the system (7.6) can be written as

$$\dot{q} = v \quad (7.11a)$$

$$\dot{v} = w \quad (7.11b)$$

$$0 = M(q)w - f(q, v) + G^T(q)\lambda \quad (7.11c)$$

$$0 = g(q). \quad (7.11d)$$

Index 2 Formulation. If we differentiate $0 = g(q)$ once and replace (7.11d) by

$$0 = G(q)v, \quad (7.11e)$$

we get an index 2 problem which is mathematically equivalent to (7.6).

Index 1 Formulation. One more differentiation of (7.11e) yields

$$0 = g_{qq}(q)(v, v) + G(q)w, \quad (7.11f)$$

so that (7.11a,b,c,f) constitutes an index 1 problem.

We have applied several codes with many different tolerances between 10^{-2} and 10^{-10} to these formulations. The results are given in Fig. 7.3. We have plotted the computing time (on a SUN Spark 20 workstation) against the error of the (q, v) -components at $x_{\text{end}} = 0.03$ (in double logarithmic scale).

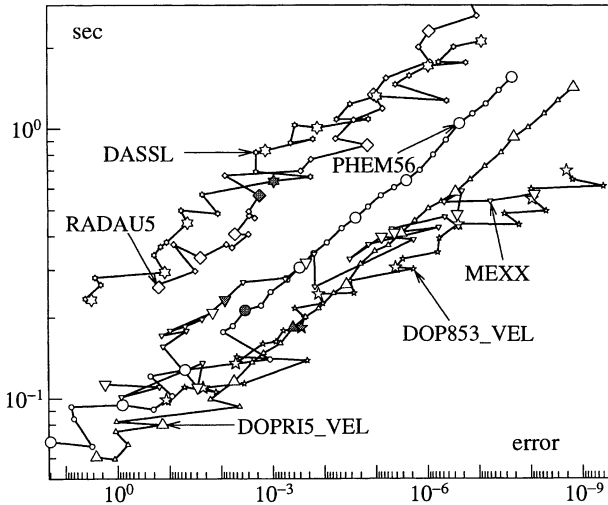


Fig. 7.3. Work-precision diagram

Explicit Runge-Kutta Methods. The index 1 formulation allows us to apply explicit methods such as DOPRI5 or DOP853 of Volume I. For this we have written a function subroutine which solves in each call the linear system (7.11c,f) for w and λ and inserts the result into (7.11a,b). Since there is no stiffness in the obtained

differential equation for (q, v) , it is not surprising that here the explicit codes work very efficiently (Fig. 7.3).

In order to avoid the drift-off phenomenon (see Sect. VII.2), we have also combined this method with projections onto the solution manifold. This can be implemented conveniently with help of the subroutine SOLOUT, which is called by DOPRI5 after every successful step (set ITRN = 2 in order to indicate that the numerical approximation has been altered). The full projection (on position and velocity level, (7.11d) and (7.11e)) is slightly more expensive than velocity stabilization alone (denoted by DOPRI5_VEL in Fig. 7.3) and does not give improved results. The first picture of Fig. 7.4 shows the results of the three different implementations: the 'standard' approach is without any projection, 'velocity' means that we perform only velocity stabilization, and 'position' indicated that we do consecutive projections on the position and velocity level. We see that velocity stabilization gives the best results concerning achieved accuracy and computing time.

Half-Explicit Methods. These methods (discussed in Sect. VII.6) are especially adapted to the numerical solution of (nonstiff) constrained mechanical systems. Only linear systems with the matrix (7.9) have to be solved, otherwise the methods are explicit. Since they are applied directly to the index 2 formulation, the velocity constraint (7.11e) is automatically satisfied, and no subroutine for the computation of $g_{qq}(q)(v, v)$ is required.

The extrapolation code MEXX of Lubich (1989) (see also Lubich, Nowak, Poehle & Engstler 1992) implements the half-explicit mid-point rule (6.18). The existence of an h^2 -expansion (Theorem 6.3) justifies extrapolation and thus yields methods of arbitrarily high order. It is not surprising that this code gives excellent results for high precision computations.

The first code implementing half-explicit Runge-Kutta methods is HEM5 of Brasey (1994). It has been modified and improved by Arnold (1995, code HEX5) and Murua (1995, code PHEM56). We have also included the results of the latter code (Fig. 7.3). It is slightly less efficient than DOPRI5_VEL in this particular example, because the evaluation of $g_{qq}(q)(v, v)$ is cheap. Arnold (1995) and Murua (1995) report about experiments (with expensive $g_{qq}(q)(v, v)$), where the half-explicit methods are superior to explicit Runge-Kutta methods with velocity projection.

BDF. The famous code DASSL of Petzold (1982), see also Brenan, Campbell & Petzold (1989), is a realization of the BDF multistep formulas. It is written for problems of the general form $F(u, u', x) = 0$, so that it is not necessary to introduce \dot{q} of (7.6) as new variable. We applied it using default values for all parameters except for the scaling of the error estimation. We put INFO(2)=1 and

$$\text{ATOL}(I) = \text{RTOL}(I) = \begin{cases} \text{Tol} & \text{for } I = 1, \dots, 14, \\ 1.0\text{D}0 & \text{for } I \geq 15, \end{cases}$$

which means that we control the accuracy for q and v , but not for the Lagrange

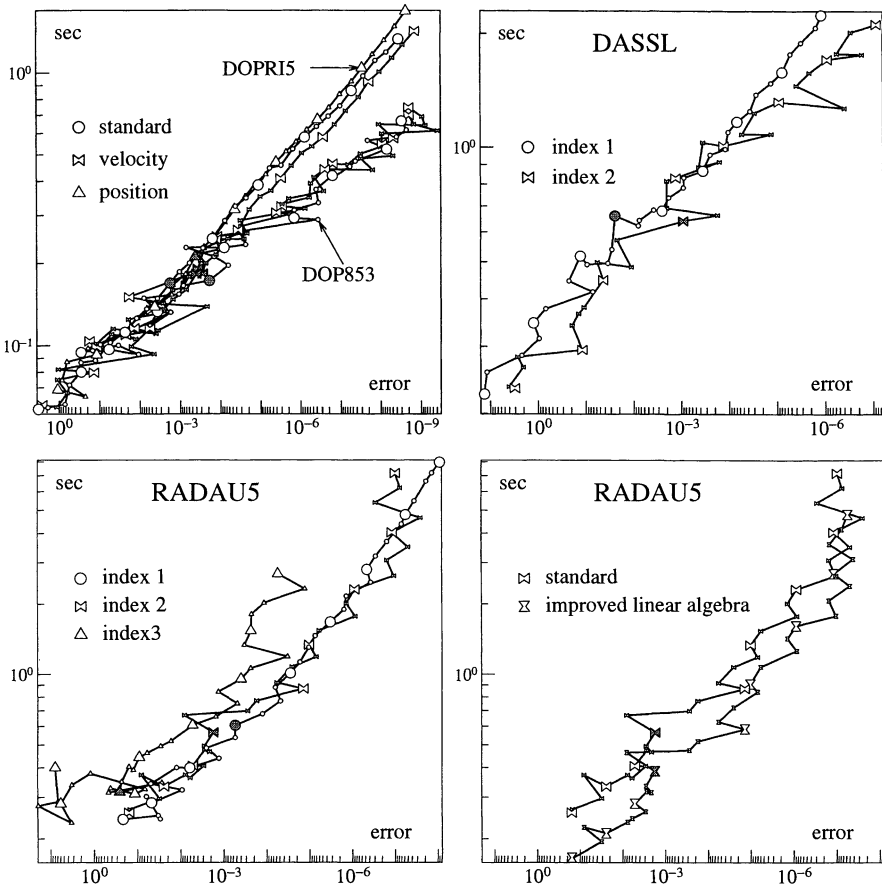


Fig. 7.4. Work-precision diagram

multipliers λ . In the comparisons of Fig. 7.3 (index 2 formulation) and Fig. 7.4 we used the full Jacobian of the problem, obtained by numerical differentiation. This turned out to be more efficient than providing an analytic approximation, where the derivatives of f , M and G are neglected.

Implicit Runge-Kutta Methods. Our code RADAU5 is written for problems of the form $By' = f(x, y)$ with constant, possibly singular matrix B . It can therefore be applied to all three of the above formulations. Convergence is guaranteed by Theorem VI.1.1 for the index 1 formulation, by Theorems 4.5 and 4.6 for the index 2 formulation, and by the results of HLR89 for the index 3 case. However, the higher the index, the more difficult is it to solve the nonlinear Runge-Kutta equations. We have applied the code with the options $IWORK(5) = 14$, $IWORK(6) = 0$ and $IWORK(7) = 13$ ($IWORK(5) = 7$ and $IWORK(6) = 7$ for the index 3 formulation), so that the acceleration w and the Lagrange multiplier λ are scaled by h^2

in the error estimation. This guarantees the convergence of the simplified Newton iterations (see HLR89, Chapter 7 for a justification). Furthermore, we have exploited the special structure $\dot{q} = v$, $\dot{v} = w$ of our system by setting $\text{IWORK}(9) = 14$ and $\text{IWORK}(10) = 7$. This speeds up the computation of the arising linear systems. The results are given in Fig. 7.3 (index 2 formulation) and in the lower left picture of Fig. 7.3 for all three formulations of the problem. We have used an analytical approximation to the Jacobian (neglecting the derivatives of f , M and G) and did not apply any projection onto the solution manifold.

Savings in Linear Algebra. If the problem is nonstiff, one can use a reduced Jacobian for the solution of the nonlinear Runge-Kutta equations. Neglecting the derivatives of f , M and G (what we have done for the above calculations), we are led to linear systems of the form (in the index 2 case)

$$\begin{pmatrix} -\alpha I & I & 0 & 0 \\ 0 & -\alpha I & I & 0 \\ 0 & 0 & M & G^T \\ 0 & G & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta q \\ \Delta v \\ \Delta w \\ \Delta \lambda \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \quad (7.12)$$

where $\alpha = (h\gamma)^{-1}$, h the step size and γ an eigenvalue of the Runge-Kutta matrix. The evaluation of the matrix in (7.12) is free, because $M(q)$ and $G(q)$ have to be evaluated anyway for the right-hand side of the differential-algebraic system. Eliminating the variable Δv in the last row of (7.12) yields the smaller system

$$\begin{pmatrix} M & G^T \\ G & 0 \end{pmatrix} \begin{pmatrix} \Delta w \\ \Delta \lambda \end{pmatrix} = \begin{pmatrix} c \\ \alpha d + Gb \end{pmatrix} \quad (7.13)$$

which is of the same type as those for the explicit methods. Once a solution to (7.13) is known the values of Δv and Δq are easily obtained from the first two rows of (7.12). We observe that the matrix in (7.13) does not depend on $\alpha = (h\gamma)^{-1}$. Hence only *one* LU decomposition is necessary for a step, independently of the number of distinct eigenvalues of the Runge-Kutta matrix. An implementation of these ideas reduced considerably the work for solving the nonlinear systems (see last picture of Fig. 7.4).

A similar reduction of the linear algebra was first proposed by Gear, Gupta & Leimkuhler (1985) for the BDF schemes. The above idea is not restricted to the index 2 case, and extends straightforwardly to the index 1 and index 3 situations. We finally remark that one has the possibility of retaining the decomposed matrix of (7.13) over several steps even in the case when the step size is changed.

A Stiff Mechanical System

We now want to introduce some “stiffness” into the above mechanical system. To this end we take into account the elasticity of one of these bodies (K_6 appears to be the simplest one) and replace it by a spring with very large spring constant c_1 . Thus the length of this spring will become an additional unknown variable q_8 . We let the unstretched length be zf (of Table 7.1), and assume that the centre of gravity C_6 has constant distance fa from the upper joint (see Fig. 7.1). We further simplify the problem by assuming that the inertia of this body remains constant. Obviously the algebraic constraints (7.5) remain unchanged; we only have to replace the constant zf in (7.5) by the new variable q_8 . The derivative matrix $G(q) = g'(q)$ has to be changed accordingly. It is now a 6×8 matrix.

The equations of motion for this modified problem are obtained as follows: in the *kinetic energy* (7.3) only the contribution of the 6th body (the new spring) changes, namely

$$\begin{aligned} C_6 : \quad x_6 &= xa + u \cdot \sin \varepsilon + (q_8 - fa) \cdot \cos(\Omega + \varepsilon) \\ y_6 &= ya - u \cdot \cos \varepsilon + (q_8 - fa) \cdot \sin(\Omega + \varepsilon) \\ \dot{x}_6^2 + \dot{y}_6^2 &= (q_8 - fa)^2 \cdot \dot{\Omega}^2 + 2 \cdot ((q_8 - fa)^2 - u \cdot (q_8 - fa) \cdot \sin \Omega) \cdot \dot{\Omega} \cdot \dot{\varepsilon} \\ &\quad + ((q_8 - fa)^2 - 2 \cdot u \cdot (q_8 - fa) \cdot \sin \Omega + u^2) \cdot \dot{\varepsilon}^2 \\ &\quad + 2 \cdot u \cdot \cos \Omega \cdot \dot{\varepsilon} \cdot \dot{q}_8 + \dot{q}_8^2 \\ \dot{\omega}_6 &= \dot{\Omega} + \dot{\varepsilon} \end{aligned}$$

In the *potential energy* we have to add a term which is due to the new spring. We thus get (compare (7.4))

$$U = -mom \cdot \beta + c_0 \cdot \frac{(\ell - \ell_0)^2}{2} + c_1 \cdot \frac{(q_8 - zf)^2}{2}, \quad (7.21)$$

where the spring constant c_1 of the new spring is large. The resulting system is again of the form (7.6), but with $q \in \mathbb{R}^8$. The initial values (7.7), (7.9), (7.12) for the 7 angles (7.2) are consistent for the new problem, if we require in addition

$$q_8(0) = zf, \quad \dot{q}_8(0) = 0. \quad (7.22)$$

This then implies $\ddot{q}_8(0) = 0$. For the choice $c_1 = 10^{10}$ we applied the implicit codes RADAU5 and DASSL to the above *stiff* mechanical system. The behaviour of these methods was nearly identical to that for the original problem (Fig. 7.4). So there was no need to draw another picture. Obviously, the explicit codes DOPRI5, PHEM56 and MEXX do not work any longer.

It should be remarked that for $Tol \leq 1/c_1$ the efficiency of the implicit codes suddenly decreases. This is due to the fact that the exact solution of the problem (with the *initial* values described above) is highly oscillatory with frequency $\mathcal{O}(\sqrt{c_1})$ and amplitude $\mathcal{O}(1/c_1)$ about a smooth solution. A general theory for such situations has been elaborated by Ch. Lubich (1993). For very stringent tolerances any code is forced to follow the oscillations and the step sizes become small.

Exercises

1. Consider the differential equation (so-called “Kreiss problem”)

$$y' = U^T(x) \begin{pmatrix} -1 & 0 \\ 0 & -1/\varepsilon \end{pmatrix} U(x)y, \quad U(x) = \begin{pmatrix} \cos x & \sin x \\ -\sin x & \cos x \end{pmatrix} \quad (7.23)$$

and apply the Runge-Kutta code RADAU5 to this stiff problem. You will observe that, for a fixed tolerance, the number of function evaluations increases with decreasing $\varepsilon > 0$. Then apply the method to the equivalent system

$$y' = z, \quad 0 = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix} U(x)z + U(x)y \quad (7.24)$$

and show that the number of function evaluations does not increase for $\varepsilon \rightarrow 0$.

- a) Explain this phenomenon by studying the convergence of the simplified Newton iterations.
- b) Prove that the index of the system (7.24) with $\varepsilon = 0$ is two.

VII.8 Symplectic Methods for Constrained Hamiltonian Systems

In principle, all approaches discussed in Sect. VII.2 can be employed for the numerical solution of constrained Hamiltonian systems. A disadvantage of these index reduction methods is, as we shall see below, that the symplectic structure of the flow is destroyed by the discretization.

In Sect. I.6 we have seen that the equations of motion for conservative mechanical systems can be written either in terms of position and velocity coordinates (Lagrangian formulation) or in terms of position and momentum coordinates (Hamiltonian formulation). For *constrained* mechanical systems the situation is exactly the same. In the present section we consider the Hamiltonian formulation

$$q' = H_p(p, q) \quad (8.1a)$$

$$p' = -H_q(p, q) - G^T(q)\lambda \quad (8.1b)$$

$$0 = g(q). \quad (8.1c)$$

Here, $H : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is the Hamiltonian function, H_p and H_q denote partial derivatives, $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ (with $m < n$) are the constraints, and $G(q) = g_q(q)$. If $T(q, \dot{q}) = \frac{1}{2}\dot{q}^T M(q)\dot{q}$ (with invertible $M(q)$) is the kinetic energy of a mechanical system and $U(q)$ its potential energy, we have $p = M(q)\dot{q}$ and

$$H(p, q) = \frac{1}{2}p^T M(q)^{-1}p + U(q), \quad (8.2)$$

(see Eq. (I.6.26)) in contrast to the Lagrange function, which is given by $\mathcal{L}(q, \dot{q}) = T(q, \dot{q}) - U(q)$. If $M(q) = I$ (the identity), we have $p = \dot{q}$ and both formulations, (1.46) and (8.1), are identical. If $M(q)$ depends on q , the formulation (8.1) may be numerically more advantageous than (1.46) (see Exercise 1).

Differentiating the constraint in (8.1) twice, we get

$$0 = G(q)H_p(p, q), \quad (8.3a)$$

$$0 = \frac{d}{dq} \left(G(q)H_p(p, q) \right) H_p(p, q) - G(q)H_{pp}(p, q) \left(H_q(p, q) + G^T(q)\lambda \right), \quad (8.3b)$$

and we see that λ can be expressed in terms of p and q , if

$$G(q)H_{pp}(p, q)G^T(q) \quad \text{is invertible} \quad (8.4)$$

in a neighbourhood of the considered solution. Therefore, (8.1) is a differential-algebraic system of index 3. If $H(p, q)$ is given by (8.2), condition (8.4) is the same as (1.47).

Properties of the Exact Flow

Every solution of the system (8.1) satisfies (8.1c) and (8.3a). It therefore lies on the manifold

$$\mathcal{M} = \{(p, q) \mid g(q) = 0, G(q)H_p(p, q) = 0\}. \quad (8.5)$$

Extracting λ from (8.3b) (this is possible, if (8.4) is satisfied), and inserting the resulting expression into (8.1b), yields a differential equation on the manifold \mathcal{M} . The situation here is completely analogous to that of (1.22) of Sect. VII.1.

Symplecticity. Our next aim is to extend the result of Theorem I.14.12 to constrained Hamiltonian systems. We consider the differential 2-form

$$\omega^2 = \sum_{I=1}^n dp^I \wedge dq^I \quad (8.6)$$

(p^I and q^I denote the components of the vectors p and q , respectively). The flow of the system (8.1), mapping an initial value $(p_0, q_0) \in \mathcal{M}$ onto $(p(t), q(t)) \in \mathcal{M}$, is denoted by φ_t . For a differentiable function $g : \mathcal{M} \rightarrow \mathcal{M}$ we further denote by $g^*\omega^2$ the differential 2-form, defined by

$$(g^*\omega^2)(\xi_1, \xi_2) = \omega^2(g'(p, q)\xi_1, g'(p, q)\xi_2).$$

This is formally identical to Definition I.14.11, but here we are only interested in the case where ξ_1 and ξ_2 lie in the tangent space

$$T_{(p,q)}\mathcal{M} = \left\{ (u, v) \mid G(q)v = 0, \frac{d}{dq}(G(q)H_p(p, q))v + G(q)H_{pp}(p, q)u = 0 \right\}$$

of the manifold (8.5).

Theorem 8.1. *The flow $\varphi_t : \mathcal{M} \rightarrow \mathcal{M}$ of the system (8.1) is a symplectic transformation on \mathcal{M} , i.e.,*

$$(\varphi_t^*\omega^2)(\xi_1, \xi_2) = \omega^2(\xi_1, \xi_2)$$

for all t , for all (p, q) , and for all ξ_1, ξ_2 lying in the tangent space $T_{(p,q)}\mathcal{M}$.

Proof. For $\xi \in T_{(p,q)}\mathcal{M}$ the tangent vector $\xi^t = \varphi'_t(p, q)\xi \in T_{(p(t), q(t))}\mathcal{M}$ is a solution of the variational equation

$$\begin{aligned} \delta \dot{p}^I &= - \sum_{J=1}^n \frac{\partial^2 H}{\partial q^I \partial p^J}(p, q) \cdot \delta p^J - \sum_{J=1}^n \frac{\partial^2 H}{\partial q^I \partial q^J}(p, q) \cdot \delta q^J \\ &\quad - \sum_{K=1}^m \lambda^K \sum_{J=1}^n \frac{\partial^2 g^K}{\partial q^I \partial q^J}(p, q) \cdot \delta q^J - \sum_{K=1}^m \frac{\partial g^K}{\partial q^I}(q) \cdot \delta \lambda^K \\ \delta \dot{q}^I &= \sum_{J=1}^n \frac{\partial^2 H}{\partial p^I \partial p^J}(p, q) \cdot \delta p^J + \sum_{J=1}^n \frac{\partial^2 H}{\partial p^I \partial q^J}(p, q) \cdot \delta q^J, \end{aligned}$$

where the $\delta\lambda^K$ (for $K = 1, \dots, m$) are obtained by differentiation of (8.3b). We now compute the time derivative of $\omega^2(\xi_1^t, \xi_2^t)$. The terms, not depending on λ or $\delta\lambda$, vanish by Theorem I.14.12. We therefore get

$$\begin{aligned} \frac{d}{dt}\omega^2(\xi_1^t, \xi_2^t) = & -\left(\sum_{K=1}^m \lambda^K \sum_{I,J=1}^n \frac{\partial^2 q^K(q)}{\partial q^I \partial q^J} dq^J \wedge dq^I \right. \\ & \left. + \sum_{K=1}^m d\lambda^K \wedge \left(\sum_{I=1}^n \frac{\partial g^K(q)}{\partial q^I} dq^I\right)\right)(\xi_1^t, \xi_2^t). \end{aligned} \quad (8.7)$$

Due to the symmetry of the second partial derivatives, the first expression of the right-hand side of Eq. (8.7) vanishes. The second expression also vanishes, because ξ_2^t lies in the tangent space $T_{(p(t), q(t))}\mathcal{M}$. Hence, $\omega^2(\xi_1^t, \xi_2^t)$ is constant, what proves the statement of the theorem. \square

Preservation of the Hamiltonian. Differentiation of $H(p(t), q(t))$ with respect to time yields

$$-H_p^T H_q - H_p^T G^T \lambda + H_q^T H_p,$$

with all expressions evaluated at $(p(t), q(t))$. The first term cancels with the last one, and the remaining term vanishes, because $G(q)H_p(p, q) = 0$ on the solution manifold. Consequently, the Hamiltonian function $H(p, q)$ is constant along solutions of (8.1).

First Order Symplectic Method

We shall now discuss in some detail the feasibility, the convergence, and the symplecticity of a simple first order method. The presented ideas will be useful for a better understanding of the later discussion of higher order methods.

Inspired by (II.16.54), we consider the following discretization of (8.1):

$$\hat{p}_1 = p_0 - h(H_q(\hat{p}_1, q_0) + G^T(q_0)\lambda_1) \quad (8.8a)$$

$$q_1 = q_0 + hH_p(\hat{p}_1, q_0) \quad (8.8b)$$

$$0 = g(q_1). \quad (8.8c)$$

The numerical approximation (\hat{p}_1, q_1) satisfies the constraint (8.1c), but not (8.3a). Therefore, we append the projection

$$p_1 = \hat{p}_1 - hG^T(q_1)\mu \quad (8.8d)$$

$$0 = G(q_1)H_p(p_1, q_1), \quad (8.8e)$$

so that method (8.8a-e) yields approximations that stay in the manifold \mathcal{M} of Eq. (8.5).

Existence of the Numerical Solution. We consider a slightly more general system than (8.8). If the initial values are not consistent, we replace the relations (8.8c) and (8.8e) by

$$g(q_1) = g(q_0) + hG(q_0)H_p(p_0, q_0) \quad (8.9a)$$

$$G(q_1)H_p(p_1, q_1) = G(q_0)H_p(p_0, q_0). \quad (8.9b)$$

We shall show that the nonlinear system (8.8a,b), (8.9a) has a locally unique solution. Inspired by the proof of Theorem 3.1 we write

$$g(q_1) - g(q_0) = \int_0^1 g_q(q_0 + \tau(q_1 - q_0)) d\tau \cdot (q_1 - q_0).$$

Inserting $g(q_1)$ from (8.9a) and q_1 from (8.8b) and dividing by h yields

$$G(q_0)H_p(p_0, q_0) = \int_0^1 g_q(q_0 + \tau(q_1 - q_0)) d\tau \cdot H_p(\hat{p}_1, q_0). \quad (8.10)$$

We next develop $H_p(\hat{p}_1, q_0)$ as

$$H_p(\hat{p}_1, q_0) = H_p(p_0, q_0) - h \int_0^1 H_{pp}(p_0 + \sigma(\hat{p}_1 - p_0), q_0) d\sigma (H_q(\hat{p}_1, q_0) + G^T(q_0)\lambda_1).$$

Inserting this formula into (8.10), an integration by parts shows that (8.9a) is equivalent to

$$\begin{aligned} 0 &= \int_0^1 (1 - \tau) g_{qq}(q_0 + \tau(q_1 - q_0)) d\tau \cdot (H_p(p_0, q_0), H_p(\hat{p}_1, q_0)) \quad (8.11) \\ &- \int_0^1 g_q(q_0 + \tau(q_1 - q_0)) d\tau \int_0^1 H_{pp}(p_0 + \sigma(\hat{p}_1 - p_0), q_0) d\sigma (H_q(\hat{p}_1, q_0) + G^T(q_0)\lambda_1). \end{aligned}$$

This is a linear system for λ_1 and allows us to express λ_1 smoothly in terms of \hat{p}_1, q_1 , and of the initial values p_0, q_0 . We insert the resulting expression for λ_1 into (8.8a). Hence, (8.8a,b) becomes a nonlinear system for \hat{p}_1, q_1 , which, for sufficiently small h , has a unique solution close to p_0, q_0 (Implicit Function Theorem). It is interesting to note that, for $h \rightarrow 0$, the value λ_1 from (8.11) does not converge to $\lambda(0)$, given by (8.3b), but to the solution λ_0 of

$$0 = \frac{1}{2} g_{qq}(H_p, H_p) - GH_{pp}(H_q + G^T \lambda_0).$$

Here, all functions are evaluated at the initial value (p_0, q_0) .

The existence of the solution (p_1, μ) to the system (8.8d), (8.9b) follows from the Newton-Kantorovich Theorem (Ortega & Rheinboldt 1970) with initial approximation $p_1 := \hat{p}_1$, and $\mu = 0$, or also from the Implicit Function Theorem.

We have not only shown that the system (8.8) possesses a locally unique solution, but we have also seen that the replacement of (8.8c,e) by (8.9) extends the definition of the method to arbitrary initial values (close to \mathcal{M}). We thus have

found a one-step method

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \begin{pmatrix} p_0 \\ q_0 \end{pmatrix} + h\Phi\left(\begin{pmatrix} p_0 \\ q_0 \end{pmatrix}, h\right) \quad (8.12)$$

in \mathbb{R}^{2n} , which reduces to (8.8) on the manifold \mathcal{M} . For smooth functions g and H also Φ is smooth, and the classical theory (convergence, asymptotic expansions, ...) can be applied to this method.

Convergence of Order 1. It is sufficient to show that the local error is of size $\mathcal{O}(h^2)$. The convergence then follows from Theorem II.3.6 applied to (8.12). From the above investigation on the existence of the numerical solution we know that $\hat{p}_1 = p_0 + \mathcal{O}(h)$, $q_1 = q_0 + \mathcal{O}(h)$, and $\lambda_1 = \lambda_0 + \mathcal{O}(h)$. Consequently, we have from (8.8a,b) that

$$q_1 = q(t_0 + h) + \mathcal{O}(h^2), \quad \hat{p}_1 = p(t_0 + h) - hG^T(q_0)\delta\lambda + \mathcal{O}(h^2) \quad (8.13)$$

with $\delta\lambda = \lambda_0 - \lambda(t_0)$. The disturbing term $hG^T(q_0)\delta\lambda$ is eliminated by the projection (8.8d,e). This can be seen as follows: from (8.13) and (8.8d) we know that $p_1 = p(t_0 + h) - G^T(q_0)\nu + \mathcal{O}(h^2)$, so that

$$G(q(t_0 + h))H_p(p(t_0 + h) - G^T(q_0)\nu, q(t_0 + h)) = \mathcal{O}(h^2).$$

By (8.4) and the Implicit Function Theorem this implies $\nu = \mathcal{O}(h^2)$, and the local error for both components (p and q) is of size $\mathcal{O}(h^2)$.

Symplecticity. Differentiation of the relations (8.8a,b) shows that (we use upper indices for the components)

$$\begin{aligned} d\hat{p}_1^I &= dp_0^I - h \sum_{J=1}^n \frac{\partial^2 H}{\partial q^I \partial p^J}(\hat{p}_1, q_0) d\hat{p}_1^J - h \sum_{J=1}^n \frac{\partial^2 H}{\partial q^I \partial q^J}(\hat{p}_1, q_0) dq_0^J \\ &\quad - h \sum_{K=1}^m \lambda_1^K \sum_{J=1}^n \frac{\partial^2 g^K}{\partial q^I \partial q^J}(q_0) dq_0^J - h \sum_{K=1}^m \frac{\partial g^K}{\partial q^I}(q_0) d\lambda_1^K \\ dq_1^I &= dq_0^I + h \sum_{J=1}^n \frac{\partial^2 H}{\partial p^I \partial p^J}(\hat{p}_1, q_0) d\hat{p}_1^J + h \sum_{J=1}^n \frac{\partial^2 H}{\partial p^I \partial q^J}(\hat{p}_1, q_0) dq_0^J. \end{aligned}$$

Taking the exterior product of the first formula with dq_0^I , and of the second formula with $d\hat{p}_1^I$, several terms cancel out (as in the proof of Theorem 8.1) and we obtain

$$\begin{aligned} \sum_{I=1}^n d\hat{p}_1^I \wedge dq_0^I &= \sum_{I=1}^n dp_0^I \wedge dq_0^I - h \sum_{I,J=1}^n \frac{\partial^2 H}{\partial q^I \partial p^J}(\hat{p}_1, q_0) d\hat{p}_1^J \wedge dq_0^I \\ \sum_{I=1}^n d\hat{p}_1^I \wedge dq_1^I &= \sum_{I=1}^n d\hat{p}_1^I \wedge dq_0^I + h \sum_{I,J=1}^n \frac{\partial^2 H}{\partial p^I \partial q^J}(\hat{p}_1, q_0) d\hat{p}_1^I \wedge dq_0^J. \end{aligned}$$

Summing up both formulas yields

$$\sum_{I=1}^n d\hat{p}_1^I \wedge dq_1^I = \sum_{I=1}^n dp_0^I \wedge dq_0^I, \quad (8.14)$$

what proves that the method (8.8a-c) is symplectic. In order to show that also the projection (8.8d,e) is symplectic, we compute

$$dp_1^I = d\hat{p}_1^I - h \sum_{K=1}^m \mu^K \sum_{J=1}^n \frac{\partial^2 g^K}{\partial q^I \partial q^J}(q_1) dq_1^J - h \sum_{K=1}^m \frac{\partial g^K}{\partial q^I}(q_1) d\mu^K,$$

and we obtain as above (using $g(q_1) = 0$) that

$$\sum_{I=1}^n dp_1^I \wedge dq_1^I = \sum_{I=1}^n d\hat{p}_1^I \wedge dq_1^I. \quad (8.15)$$

Equations (8.14) and (8.15) together show that the complete procedure (8.8a-e) is symplectic.

SHAKE and RATTLE

These algorithms have been designed for problems with separable Hamiltonian

$$H(p, q) = \frac{1}{2} p^T M^{-1} p + U(q) \quad (8.16)$$

(constant matrix M), and are very popular in molecular dynamics simulation. Observe that for this Hamiltonian the problem (8.1) becomes the second order differential equation $Mq'' = -U_q(q) - G^T(q)\lambda$ with constraint (8.1c).

SHAKE. This method, due to Ryckaert, Ciccotti & Berendsen (1977), is given by

$$q_{n+1} - 2q_n + q_{n-1} = -h^2 M^{-1} (U_q(q_n) + G^T(q_n)\lambda_n) \quad (8.17a)$$

$$0 = g(q_{n+1}). \quad (8.17b)$$

In the absence of constraints it is identical to Störmer's method (Sect. III.10), which in molecular dynamics applications is often referred the Verlet method (Verlet 1967). The p -components are approximated by $p_n = M(q_{n+1} - q_{n-1})/2h$. For an implementation of this 2-step method a stabilized version is recommended (see the end of Sect. III.10).

RATTLE. Denoting $p_{n+1/2} := p_n - (h/2)(U_q(q_n) + G^T(q_n)\lambda_n)$, the SHAKE algorithm can be rewritten in the form

$$p_{n+1/2} = p_n - \frac{h}{2} (U_q(q_n) + G^T(q_n)\lambda_n) \quad (8.18a)$$

$$q_{n+1} = q_n + h M^{-1} p_{n+1/2} \quad (8.18b)$$

$$0 = g(q_{n+1}). \quad (8.18c)$$

The definition of p_{n+1} as in the SHAKE method requires the knowledge of q_{n+2} . In order to avoid this difficulty, Andersen (1983) suggests to define p_{n+1} by

$$p_{n+1} = p_{n+1/2} - \frac{h}{2}(U_q(q_{n+1}) + G^T(q_{n+1})\mu_n) \quad (8.18d)$$

$$0 = G(q_{n+1})M^{-1}p_{n+1}, \quad (8.18e)$$

so that also the hidden constraint (8.3a) is satisfied. These two equations constitute a linear system for (p_{n+1}, μ_n) .

Extension to General Hamiltonian Functions. It was observed by Jay (1994) that the RATTLE algorithm can be extended to general Hamiltonian functions as follows: for consistent values $(p_n, q_n) \in \mathcal{M}$ define

$$p_{n+1/2} = p_n - \frac{h}{2}(H_q(p_{n+1/2}, q_n) + G^T(q_n)\lambda_n) \quad (8.19a)$$

$$q_{n+1} = q_n + \frac{h}{2}(H_p(p_{n+1/2}, q_n) + H_p(p_{n+1/2}, q_{n+1})) \quad (8.19b)$$

$$0 = g(q_{n+1}). \quad (8.19c)$$

$$p_{n+1} = p_{n+1/2} - \frac{h}{2}(H_q(p_{n+1/2}, q_{n+1}) + G^T(q_{n+1})\mu_n) \quad (8.19d)$$

$$0 = G(q_{n+1})H_p(p_{n+1}, q_{n+1}). \quad (8.19e)$$

This is the special case $s = 2$ of the Lobatto IIIA-IIIb pair to be discussed below.

The equations (8.19a-c) constitute a nonlinear system for the unknowns $p_{n+1/2}$, q_{n+1} , and λ_n . In the same way as for the method (8.8) we can reformulate Eq. (8.19c) in such a way that λ_n can be expressed smoothly in terms of p_n , q_n , $p_{n+1/2}$, q_{n+1} , and h . Hence, the numerical solution exists, is locally unique, and depends smoothly on h and on the initial values (p_n, q_n) . The same is true for the system (8.19d,e). If the equations (8.19c,e) are replaced by (8.9), we get a smooth extension of the method (8.19), defined on a neighbourhood of \mathcal{M} in \mathbb{R}^{2n} .

Theorem 8.2. *The numerical method (8.19) is symmetric, convergent of order 2, and symplectic.*

Proof. a) We consider the more general situation, where (8.19c,e) is replaced by (8.9). Replacing then h by $-h$, and exchanging (p_n, q_n) with (p_{n+1}, q_{n+1}) and

λ_n with μ_n , we obtain

$$\begin{aligned} p_{n+1/2} &= p_{n+1} + \frac{h}{2} \left(H_q(p_{n+1/2}, q_{n+1}) + G^T(q_{n+1})\mu_n \right) \\ q_n &= q_{n+1} - \frac{h}{2} \left(H_p(p_{n+1/2}, q_{n+1}) + H_p(p_{n+1/2}, q_n) \right) \\ g(q_n) &= g(q_{n+1}) - hG(q_{n+1})H_p(p_{n+1}, q_{n+1}) \\ p_n &= p_{n+1/2} + \frac{h}{2} \left(H_q(p_{n+1/2}, q_n) + G^T(q_n)\lambda_n \right) \\ G(q_n)H_p(p_n, q_n) &= G(q_{n+1})H_p(p_{n+1}, q_{n+1}). \end{aligned}$$

These are exactly the same equations as those of (8.19a,b,d) and (8.9), proving that even the extension of the method to a neighbourhood of \mathcal{M} is symmetric.

b) We consider the method (8.19) as a mapping $(p_n, q_n) \mapsto (p_{n+1}, q_{n+1})$ on the manifold \mathcal{M} of Eq. (8.5). The same considerations as for (8.8) show that (8.19) is a method of order at least one. Since it is symmetric, its order has to be even (Sect. II.8). This proves that (8.19) is a convergent method of order 2.

c) The fact that the method (8.19) defines a symplectic transformation on \mathcal{M} can be proved as for (8.8) (see Leimkuhler & Skeel (1994) for the case of a separable Hamiltonian (8.16)). We do not give details here, because the symplecticity of (8.19) also follows from Theorem 8.5 below. \square

Remark 8.3. In a step by step application of method (8.19) the projection (8.19d,e) can be avoided at those points, where the value p_{n+1} is not needed for output. Indeed, from the second step on we can replace (8.19a) by

$$p_{n+1/2} = p_{n-1/2} - \frac{h}{2} \left(H_q(p_{n+1/2}, q_n) + H_q(p_{n-1/2}, q_n) + G^T(q_n)(\lambda_n + \mu_{n-1}) \right)$$

without changing the numerical approximations q_n and $p_{n+1/2}$. The same trick is possible for method (8.8).

The Lobatto IIIA-IIIB Pair

Partitioned Runge-Kutta methods are well suited for unconstrained Hamiltonian systems (see Sect. II.16). We shall investigate here, how these methods can be extended to the constrained system (8.1). We consider

$$P_i = p_0 + h \sum_{j=1}^s a_{ij} k_j, \quad Q_i = q_0 + h \sum_{j=1}^s \hat{a}_{ij} \ell_j, \quad (8.20a)$$

$$p_1 = p_0 + h \sum_{i=1}^s b_i k_i, \quad q_1 = q_0 + h \sum_{i=1}^s \hat{b}_i \ell_i, \quad (8.20b)$$

$$k_i = -\frac{\partial H}{\partial q}(P_i, Q_i) - G^T(Q_i)\Lambda_i, \quad \ell_i = \frac{\partial H}{\partial p}(P_i, Q_i), \quad (8.20c)$$

where b_i, a_{ij} and $\widehat{b}_i, \widehat{a}_{ij}$ are the coefficients of two Runge-Kutta schemes (c.f., Eq. (II.16.26)). For the moment, the values Λ_i ($i = 1, \dots, s$) are not yet specified. There are several possibilities to do this. One can either define them by $\Lambda_i = \lambda(P_i, Q_i)$, where $\lambda(p, q)$ is the function given by (8.3b), or one can define them implicitly by adding the conditions $G(Q_i)H_p(P_i, Q_i) = 0$ or $g(Q_i) = 0$.

We are interested in symplectic schemes. Therefore it is natural to consider methods satisfying the conditions of Theorem II.16.10.

Lemma 8.4. *If the coefficients of (8.20) satisfy*

$$b_i = \widehat{b}_i, \quad i = 1, \dots, s \quad (8.21)$$

$$b_i \widehat{a}_{ij} + \widehat{b}_j a_{ji} - b_i \widehat{b}_j = 0, \quad i, j = 1, \dots, s, \quad (8.22)$$

then we have the following relation for the expressions in (8.20):

$$\sum_{I=1}^n dp_1^I \wedge dq_1^I - \sum_{I=1}^n dp_0^I \wedge dq_0^I = h \sum_{i=1}^s b_i \sum_{K=1}^m \left(\sum_{I=1}^n \frac{\partial g^K}{\partial q^I}(Q_i) dQ_i^I \right) \wedge d\Lambda_i^K.$$

If the Hamiltonian is separable (i.e., $H(p, q) = T(p) + U(q)$), then the condition (8.22) alone implies the above relation.

Proof. We compute the expression $D = \sum_I dp_1^I \wedge dq_1^I - \sum_I dp_0^I \wedge dq_0^I$ following the lines of the proof of Theorem II.16.6 (see also the proof of Theorem II.16.10). All terms cancel with exception of those originating from the presence of $G^T(Q_i)\Lambda_i$ in (8.20c). We thus obtain

$$D = -h \sum_{i=1}^s b_i \sum_{K=1}^m \left(\Lambda_i^K \sum_{I, J=1}^n \frac{\partial^2 g^K}{\partial q^J \partial q^I}(Q_i) dQ_i^J \wedge dQ_i^I + \sum_{I=1}^n \frac{\partial g^K}{\partial q^I}(Q_i) d\Lambda_i^K \wedge dQ_i^I \right).$$

Due to the symmetry of the second derivative of g^K the term involving $dQ_i^J \wedge dQ_i^I$ vanishes identically. This proves the statement of the lemma. \square

We are interested in partitioned Runge-Kutta methods that satisfy:

- the numerical solution stays on the manifold \mathcal{M} of Eq. (8.5);
- the numerical flow $(p_0, q_0) \mapsto (p_1, q_1)$ is a symplectic transformation on \mathcal{M} ;
- the order of convergence is higher than 2.

If the values Λ_i are determined by the condition

$$g(Q_i) = 0 \quad \text{for } i = 1, \dots, s, \quad (8.23)$$

then we have $\sum_I \partial g^K / \partial q^I(Q_i) dQ_i^I = 0$, and it follows from Lemma 8.4 that the method (8.20) is symplectic, if (8.21) and (8.22) are satisfied. Hence, the second item holds. Here we see the importance of the conditions (8.23). Solving the index reduced system (8.1a,b), (8.3b) by a symplectic method would in general not result in a symplectic numerical flow on \mathcal{M} .

How can we achieve the first item, in particular the condition $g(q_1) = 0$? The idea is to require the method \hat{b}_i, \hat{a}_{ij} to be stiffly accurate, i.e.,

$$\hat{a}_{sj} = \hat{b}_j \quad \text{for } j = 1, \dots, s. \quad (8.24)$$

In this case we have $q_1 = Q_s$, and $g(q_1) = 0$ is automatically satisfied by (8.23). The condition (8.24) together with (8.22) implies that (assuming nonzero \hat{b}_i)

$$a_{is} = 0 \quad \text{for } i = 1, \dots, s, \quad (8.25)$$

and the nonlinear system (8.20a,c), (8.23) no longer depends on Λ_s . This parameter, however, appears in the definition of p_1 in Eq. (8.20b) via k_s . There it can be used to impose the constraint $G(q_1)H_p(p_1, q_1) = 0$.

Due to the condition (8.25) a new difficulty arises. If we consider (8.20b,c) as definition of the quantities p_1, q_1, k_i, ℓ_i , the remaining equations (8.20a) and (8.23) are a nonlinear system for $P_1, \dots, P_s, Q_1, \dots, Q_s, \Lambda_1, \dots, \Lambda_{s-1}$. Counting the number of equations of this system ($2sn + sm$) and the number of unknowns ($2sn + (s-1)m$), one is readily convinced that this nonlinear system will usually not have a solution. The idea (Jay 1994, 1996) is to require

$$\hat{a}_{1j} = 0 \quad \text{for } j = 1, \dots, s, \quad (8.26)$$

so that $Q_1 = q_0$, and the condition (8.23) is automatically verified for $i = 1$ (we always assume consistent initial values). By (8.22) this implies (for nonzero \hat{b}_i)

$$a_{i1} = b_1 \quad \text{for } i = 1, \dots, s. \quad (8.27)$$

The Runge-Kutta matrices \hat{A} and A are both singular. Let \hat{A}_0 be the $(s-1) \times s$ submatrix of \hat{A} obtained by deleting its first row, and let A_0 be the $s \times (s-1)$ submatrix of A formed by the first $s-1$ columns of A . In order to be able to prove the existence of a numerical solution of (8.20), (8.23), we require that the $(s-1) \times (s-1)$ matrix

$$\hat{A}_0 A_0 \quad \text{is invertible.} \quad (8.28)$$

We now extend the method to arbitrary initial values as follows: we replace condition (8.23) by

$$g(Q_i) = g(q_0) + \hat{c}_i h G(q_0) H_p(p_0, q_0) \quad \text{for } i = 1, \dots, s,$$

($\hat{c}_i = \sum_j \hat{a}_{ij}$) and the condition $G(q_1)H_p(p_1, q_1) = 0$ by (8.9b). Similar to Equation (8.10) we use

$$g(Q_i) - g(q_0) = h \int_0^1 g_q(q_0 + \tau(Q_i - q_0)) d\tau \cdot \sum_{j=1}^s \hat{a}_{ij} H_p(P_j, Q_j). \quad (8.29)$$

Then we develop

$$\begin{aligned} H_p(P_j, Q_j) &= H_p(p_0, Q_j) \\ &- h \int_0^1 H_{pp}(p_0 + \sigma(P_j - p_0), Q_j) d\sigma \cdot \sum_{r=1}^{s-1} a_{jr} (H_q(P_r, Q_r) + G^T(Q_r) \Lambda_r), \end{aligned}$$

and insert this relation into (8.29). As in Eq. (8.11) we get a linear system for $\Lambda_1, \dots, \Lambda_{s-1}$ which, for $h = 0$, has the solution Λ_r^0 given by

$$0 = \frac{\widehat{c}_i^2}{2} g_{qq}(H_p, H_p) + \left(\sum_{j=1}^s \widehat{a}_{ij} \widehat{c}_j \right) G H_{pq} H_p \\ - \sum_{r=1}^{s-1} \left(\sum_{j=1}^s \widehat{a}_{ij} a_{jr} \right) G H_{pp} (H_q + G^T \Lambda_r^0).$$

Here all functions are evaluated at (p_0, q_0) . Due to (8.28) and (8.4) this system can be solved for Λ_r^0 . The Implicit Function Theorem then guarantees the existence of a locally unique solution of the method (8.20), (8.23), and the existence of a smooth extension to a neighbourhood of \mathcal{M} .

The question is now: do there exist high order methods having all these properties?

Theorem 8.5. *The s -stage Lobatto IIIA-IIIB pair (Lobatto IIIA in the role of $\widehat{b}_i, \widehat{a}_{ij}$, and Lobatto IIIB in the role of b_i, a_{ij} ; see Sect. IV.5 for their definition) satisfies (8.21), (8.22), (8.24), (8.25), (8.26), (8.27), and (8.28).*

Proof. Properties (8.21), (8.24), (8.25), (8.26), and (8.27) follow immediately from the definition of the methods. The symplecticity condition (8.22) has first been proved by Sun Geng (1993). We let $d_{ij} = b_i \widehat{a}_{ij} + \widehat{b}_j a_{ji} - b_i \widehat{b}_j$ and compute for $k = 1, \dots, s$

$$\sum_{j=1}^s d_{ij} c_j^{k-1} = b_i \frac{c_i^k}{k} + \frac{b_i}{k} (1 - c_i^k) - b_i \frac{1}{k} = 0.$$

Here we have exploited the fact that the Lobatto IIIA method satisfies $C(s)$ and the Lobatto IIIB method satisfies $D(s)$ (see Table IV.5.13). Since the abscissae c_1, \dots, c_s of the Lobatto quadrature are distinct, the above Vandermonde type system has a unique solution $d_{ij} = 0$. This proves (8.22).

We next show that

$$\sum_{k=1}^{s-1} \left(\sum_{j=1}^s \widehat{a}_{ij} a_{jk} \right) c_k^{q-2} = \frac{c_i^q}{q(q-1)} \quad \text{for } i, q = 2, \dots, s. \quad (8.30)$$

This means that $\widehat{A}_0 A_0 V = W$, where V and W are nonsingular Vandermonde type matrices. This obviously implies (8.28). For $q = 2, \dots, s-1$ Eq. (8.30) follows from the fact that the methods Lobatto IIIA and IIIB satisfy $C(s)$ and $C(s-2)$, respectively. It remains to show that the coefficients $\delta_i := \sum_k \sum_j \widehat{a}_{ij} a_{jk} c_k^{s-2} - c_i^s / (s-1)$ vanish for all i . By (8.26) and $c_1 = 0$ we have $\delta_1 = 0$. Because of $\widehat{a}_{sj} = \widehat{b}_j = b_j$ and $c_s = 1$, the condition $\delta_s = 0$ is nothing else than an order condition (order s), which is satisfied (Sect. IV.5). Since the Lobatto IIIA and IIIB methods satisfy $D(s-2)$ and $D(s)$, respectively, it holds $\sum_i b_i c_i^{m-1} \delta_i = 0$ for

$m = 1, \dots, s-2$. This proves that also $\delta_2, \dots, \delta_{s-1}$ vanish, so that all relations of (8.30) are established. \square

It still remains to discuss the order of convergence of the Lobatto IIIA-IIIB pair. Since we have succeeded in embedding the method into a one-step method that is defined in a whole neighbourhood of \mathcal{M} , the convergence theory of Sect. II.3 can be applied. We only have to investigate the local error of the method. Each of the methods has classical order $2s-2$ (Sect. IV.5), and it follows from Exercise 4 that, considered as partitioned Runge-Kutta method, the pair has also order $2s-2$. It has been shown in Jay (1994) that the presence of constraints (8.1c) does not reduce the order. The proof of this superconvergence result is very technical and long. Therefore we do not reproduce it here.

Composition Methods

Another possibility for obtaining high order symplectic methods for the system (8.1) is by composition of low order methods. The idea goes back to Yoshida (1990), and has been extended to constrained systems by Reich (1996).

Consider the second order symmetric method (8.19) and denote its extension to a neighbourhood of \mathcal{M} by Φ_h . We shall study the following composition

$$\Phi_{c_1 h} \circ \Phi_{c_2 h} \circ \Phi_{c_1 h}. \quad (8.31)$$

The method (8.31) represents a one-step method, defined in a neighbourhood of \mathcal{M} . For initial values on \mathcal{M} , the numerical solution stays on \mathcal{M} . Moreover, the composition (8.31) is symplectic and symmetric. Observe that the projections (8.19d,e) can be avoided in an implementation of this method (see Remark 8.3). Concerning its order we have the following result.

Theorem 8.6. *Let Φ_h be the mapping $(p_0, q_0) \mapsto (p_1, q_1)$, defined by (8.19). If*

$$2c_1 + c_2 = 1, \quad 2c_1^3 + c_2^3 = 0, \quad (8.32)$$

the composition method (8.31) is of order 4.

If Φ_h represents a one-step method that is symmetric, of order $p = 2k$, and defined in a neighbourhood of \mathcal{M} , then the relations

$$2c_1 + c_2 = 1, \quad 2c_1^{p+1} + c_2^{p+1} = 0, \quad (8.33)$$

imply that the composition (8.31) is of order $p+2$.

Proof. We let $y_0 = (p_0, q_0)^T$ and $y(t) = (p(t), q(t))^T$. The local error of the method (8.19) satisfies

$$y(t_0 + h) - \Phi_h(y_0) = d(y_0)h^3 + \mathcal{O}(h^4).$$

Since the basic method is of the form $\Phi_h(y_0) = y_0 + h\Psi(y_0, h)$, we have that

$$y(t_0 + (2c_1 + c_2)h) - \Phi_{c_1 h} \circ \Phi_{c_2 h} \circ \Phi_{c_1 h}(y_0) = (2c_1^3 + c_2^3)d(y_0)h^3 + \mathcal{O}(h^4).$$

The conditions (8.32) then imply that the method (8.31) is at least of order 3. Since it is symmetric, it has to be of order 4. The proof is easily adapted to the higher order situation. \square

A solution of (8.32) is given by

$$c_1 = \frac{1}{2 - \sqrt[3]{2}}, \quad c_2 = -\frac{\sqrt[3]{2}}{2 - \sqrt[3]{2}},$$

which shows that the intermediate step in the composition (8.31) is a ‘back step’ (negative step size $c_2 h$).

The result of Theorem 8.6 allows us to construct symplectic integrators for (8.1) of an arbitrary even order. However, the resulting method of order $p = 2k$ requires 3^{k-1} applications of the basic method (8.19).

In the case of unconstrained Hamiltonian systems it is known that better methods can be obtained by compositions of the form

$$\Phi_{c_1 h} \circ \Phi_{c_2 h} \circ \dots \circ \Phi_{c_{s-1} h} \circ \Phi_{c_s h} \circ \Phi_{c_{s-1} h} \circ \dots \circ \Phi_{c_2 h} \circ \Phi_{c_1 h} \quad (8.34)$$

(see Yoshida 1990, McLachlan 1995, Sanz-Serna & Calvo 1994). Reich (1996) studies the extension of these methods to constrained Hamiltonian systems and finds that additional order conditions are necessary. His investigation relies on a “backward error analysis” for integrators on manifolds.

Backward Error Analysis (for ODEs)

Although backward analysis is a perfectly straightforward concept there is strong evidence that a training in classical mathematics leaves one unprepared to adopt it.

(J.H. Wilkinson, NAG Newsletter 2/85)

In Sect. II.16 we have briefly explained the idea of backward error analysis for the symplectic Euler method. Here we present an extension to general one-step methods for ordinary differential equations. Consider

$$y' = f(y), \quad y(0) = y_0, \quad (8.35)$$

and let $y_0 \mapsto y_1$ be an arbitrary one-step method for (8.35). We assume that $f(y)$ and the method are sufficiently often differentiable, so that the local error can be expanded into a Taylor series as

$$y_1 - y(h) = d_{p+1}(y_0)h^{p+1} + \dots + d_N(y_0)h^N + \mathcal{O}(h^{N+1}). \quad (8.36)$$

Theorem 8.7. *Consider a one-step method of order p , and assume the local error to be given by (8.36). Then there exist functions $f_j(y)$ (for $j = p, \dots, N$), such that*

$$y_1 - \tilde{y}(h) = \mathcal{O}(h^{N+1}), \quad (8.37)$$

where $\tilde{y}(t)$ is the solution of the perturbed differential equation

$$\tilde{y}' = f(\tilde{y}) + h^p f_p(\tilde{y}) + \dots + h^{N-1} f_{N-1}(\tilde{y}), \quad \tilde{y}(0) = y_0, \quad (8.38)$$

Remark. If the function $f(y) + h^p f_p(y) + \dots + h^{N-1} f_{N-1}(y)$ satisfies a Lipschitz condition, the proof of Theorem II.3.4 shows that $y_n - \tilde{y}(nh) = \mathcal{O}(h^N)$ on bounded intervals. This implies that the numerical approximation y_n is much closer to the solution of (8.38) than to that of (8.35). Hence, the study of the system (8.38) yields new insight into the behaviour of the numerical solution.

Proof. As a consequence of the nonlinear variation-of-constants formula (Theorem I.14.5) we have

$$\tilde{y}(h) = y(h) + \int_0^h \frac{\partial y}{\partial y_0}(h, s, \tilde{y}(s)) \cdot \left(h^p f_p(\tilde{y}(s)) + \dots + h^N f_N(\tilde{y}(s)) \right) ds,$$

where $y(t, t_0, y_0)$ denotes the solution of (8.35) corresponding to initial values $y(t_0) = y_0$. Expanding the above integral into a Taylor series we obtain

$$\tilde{y}(h) - y(h) = h^{p+1} f_p(y_0) + h^{p+2} \left(f_{p+1} + \frac{1}{2} f'_p f + \frac{1}{2} f' f'_p \right)(y_0) + \dots \quad (8.39)$$

The condition (8.37) implies that the coefficients of (8.39) have to agree with those of (8.36) up to a certain order. We thus get $f_p(y) = d_{p+1}(y)$, $f_{p+1}(y) = d_{p+2}(y) - (f'_p(y)f(y) + f'(y)f_p(y))/2$, etc. The essential observation is that the coefficient of h^{j+1} in (8.39) contains $f_j(y)$ as linear term and further expressions that only depend on $f_i(y)$ with $i < j$. Hence, the functions $f_j(y)$ are recursively determined by the above comparison. \square

Example 8.8. For an illustration of the above theorem we consider the Volterra-Lotka differential equation

$$u' = u(v-1), \quad v' = v(2-u). \quad (8.40)$$

This system possesses the first integral

$$I(u, v) = 2 \ln u - u + \ln v - v, \quad (8.41)$$

implying that the solutions are all periodic. Some of them are plotted in the left upper picture of Fig. 8.1.

We apply three different numerical methods to this differential equation. The first one is the well-known explicit Euler method $y_{n+1} = y_n + hf(y_n)$. The right upper picture of Fig. 8.1 shows the numerical solution and the exact solution (solid

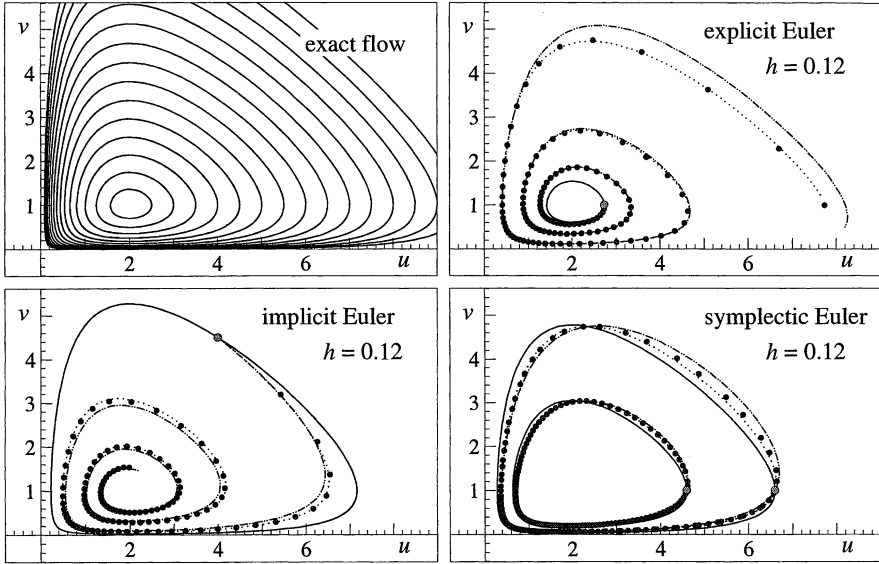


Fig. 8.1. Solutions of the perturbed differential equation for various methods

line) for the initial value $u_0 = 2.725$, $v_0 = 1$. Moreover, we have included the solutions of the perturbed differential equation (8.38) for $N = 1$ (dashed-dotted line) and for $N = 2$ (dotted line). For the explicit Euler method, Eq. (8.38) reads

$$\tilde{y}' = f(\tilde{y}) - \frac{h}{2}(f'f)(\tilde{y}) + \frac{h^2}{12}(f''(f, f) + 4f'f'f)(\tilde{y}). \quad (8.42)$$

We nicely observe the good agreement of the numerical solution with the exact solution of the perturbed system, even for the rather large step size $h = 0.12$.

The left lower picture shows the same experiment for the implicit Euler method $y_{n+1} = y_n + hf(y_{n+1})$. The perturbed differential equation is obtained from (8.42) by replacing h by $-h$ (this is, because the explicit Euler method is the adjoint method of the implicit Euler method).

The third method is the symplectic Euler method (see Eq. (8.45) below), which for the problem (8.40) is defined by

$$u_{n+1} = u_n + hu_n(v_{n+1} - 1), \quad v_{n+1} = v_n + hv_{n+1}(2 - u_n).$$

The first term of the perturbed differential equation is

$$\begin{aligned} \tilde{u}' &= \tilde{u}(\tilde{v} - 1) - h\tilde{u}(\tilde{u}\tilde{v} - 4\tilde{v} + \tilde{v}^2 + 1)/2 \\ \tilde{v}' &= \tilde{v}(2 - \tilde{u}) + h\tilde{v}(\tilde{u}\tilde{v} - 5\tilde{u} + \tilde{u}^2 + 4)/2. \end{aligned} \quad (8.43)$$

The qualitative behaviour of this method is quite different from that of the previous methods. One can prove that the system (8.43) has a first integral close to $I(u, v)$ (Exercise 5). Hence the solutions are periodic, as it is the case for the original unperturbed system.

Example 8.9. For the Hamiltonian system (without constraints)

$$q' = H_p(p, q), \quad p' = -H_q(p, q) \quad (8.44)$$

the method (8.8) becomes

$$q_1 = q_0 + hH_p(p_1, q_0), \quad p_1 = p_0 - hH_q(p_1, q_0). \quad (8.45)$$

A similar method (implicit in q and explicit in p) has been considered in Sect. II.16, Formula (II.16.54). There we have computed the first terms of the perturbed differential equation (8.38), and we have noticed with surprise that it is also Hamiltonian. The same computation can be done here. We find that the perturbed differential equation for (8.45) is of the form

$$\tilde{q}' = \tilde{H}_p(\tilde{p}, \tilde{q}), \quad \tilde{p}' = -\tilde{H}_q(\tilde{p}, \tilde{q}) \quad (8.46)$$

with (for $N = 2$)

$$\tilde{H} = H - \frac{h}{2} H_p H_q + \frac{h^2}{12} (H_{pp} H_q^2 + H_{qq} H_p^2 + 4H_{pq} H_p H_q).$$

For notational convenience we have assumed that p and q are scalars. However, with a suitable interpretation of the appearing expressions, the formula is also valid for problems with more than one degree of freedom.

Example 8.10. The second order method (8.19), when applied to the unconstrained system (8.44), becomes

$$\begin{aligned} q_1 &= q_0 + \frac{h}{2} (H_p(p_{1/2}, q_0) + H_p(p_{1/2}, q_1)) \\ p_1 &= p_0 - \frac{h}{2} (H_q(p_{1/2}, q_0) + H_q(p_{1/2}, q_1)), \end{aligned} \quad (8.47)$$

where $p_{1/2} = p_0 - (h/2)H_q(p_{1/2}, q_0)$. Computing the dominant term of its local error, we see that the perturbed differential equation (8.38) is, for $N = 2$, given by

$$\begin{aligned} \tilde{q}' &= H_p(\tilde{p}, \tilde{q}) + \frac{h^2}{24} (-H_{ppp} H_q^2 + 2H_{ppq} H_p H_q + 2H_{pqq} H_p^2 \\ &\quad + 2H_{pq} H_{pq} H_p + 4H_{pp} H_{qq} H_p) (\tilde{p}, \tilde{q}) \\ \tilde{p}' &= -H_q(\tilde{p}, \tilde{q}) + \frac{h^2}{24} (H_{ppq} H_q^2 - 2H_{pqq} H_p H_q - 2H_{qqq} H_p^2 \\ &\quad - 2H_{pq} H_{pq} H_q + 2H_{pp} H_{qq} H_q - 6H_{pq} H_{qq} H_p) (\tilde{p}, \tilde{q}). \end{aligned}$$

One easily verifies that this is a Hamiltonian system (8.46) with

$$\tilde{H} = H + \frac{h^2}{24} (2H_{qq} H_p^2 - H_{pp} H_q^2 + 2H_{pq} H_p H_q).$$

A Short Survey on Further Results. A further elaboration of backward error analysis for ordinary differential equations would take us beyond the scope of this chapter. We therefore collect some interesting results without going into details.

First of all, the mystery of the foregoing examples is well understood. In the situation, where the differential equation (8.35) is a Hamiltonian system, and where a symplectic integration method is applied, the perturbed system (8.38) is again Hamiltonian for all N . This result is proved by Hairer (1994), where explicit formulas for the functions $f_j(y)$ in terms of elementary differentials are provided, and where an explicit formula for the perturbed Hamiltonian is given. This explicit representation guarantees that $\tilde{H}(p, q)$ is uniquely defined on regions where $H(p, q)$ is defined. Different proofs of this result can be found in Reich (1996) and Benettin & Giorgilli (1994).

If the function f in (8.35) is infinitely differentiable, then the truncation index N in Theorem 8.7 is arbitrary. In general, the series (8.38) diverges as $N \rightarrow \infty$ and the constants hidden in the $\mathcal{O}(h^{N+1})$ bounds of (8.37) tend to infinity with N , even if f is analytic. Therefore, it is interesting to find rigorous bounds on $y_1 - \tilde{y}(h)$ for an optimally chosen N . Such results have been found independently by Benettin & Giorgilli (1994) and Hairer & Lubich (1996). As a consequence, one can show that for symplectic integrations the Hamiltonian remains bounded (with error of size $\mathcal{O}(h^p)$) over exponentially long times. Moreover, KAM theory can be applied to get more insight into the long-time behaviour of symplectic numerical schemes.

Backward Error Analysis on Manifolds

Consider the constrained Hamiltonian system (8.1), and a numerical one-step method which yields approximations (p_n, q_n) staying on the manifold \mathcal{M} of Eq. (8.5). Can we extend the above backward error analysis for ODEs to this situation?

There are at least two ways to achieve this goal. The first one is to introduce local coordinates in order to obtain an unconstrained Hamiltonian system. The backward analysis for ODEs can then be applied to the one-step method written in local coordinates.

The second approach allows us to construct the perturbed Hamiltonian directly in the original coordinates. For the special case of separable Hamiltonians, this approach is due to Reich (1996). We shall explain it for the first and second order methods (8.8) and (8.19).

Backward Error Analysis for the Method (8.8). Consider first the subsystem (8.8a-c). The projection step (8.8d,e) will be treated later. In Eq. (8.11) the value λ_1 has been expressed in terms of \hat{p}_1, q_1, p_0, q_0 , even for inconsistent initial values. Inserting this function into (8.8a), the Eqs. (8.8a,b) represent two relations between the variable \hat{p}_1, q_1, p_0, q_0 , and h . By the Implicit Function Theorem these two

relations allow us to express (p_0, q_1) in terms of (\hat{p}_1, q_0) , and h . Consequently, the solution λ_1 of Eq. (8.11) can be written as a function of (\hat{p}_1, q_0, h) . We denote it by

$$\lambda_1 = \lambda(\hat{p}_1, q_0, h), \quad (8.48)$$

so that the system (8.8a,b) becomes

$$\begin{aligned} \hat{p}_1 &= p_0 - h(H_q(\hat{p}_1, q_0) + G^T(q_0)\lambda(\hat{p}_1, q_0, h)) \\ q_1 &= q_0 + hH_p(\hat{p}_1, q_0), \end{aligned} \quad (8.49)$$

and the constraint (8.9a) is automatically satisfied by the definition of $\lambda(\hat{p}_1, q_0, h)$. We now consider the Hamiltonian function

$$\mathcal{H}(p, q) = H(p, q) + g(q)^T \lambda(p, q, h), \quad (8.50)$$

where $\lambda(p, q, h)$ is the function defined in (8.48). The corresponding Hamiltonian system is

$$\begin{aligned} q' &= H_p(p, q) + g(q)^T \lambda_p(p, q, h) \\ p' &= -H_q(p, q) - G^T(q)\lambda(p, q, h) - g(q)^T \lambda_q(p, q, h). \end{aligned} \quad (8.51)$$

The main observation is now that, for initial values satisfying $g(q_0) = 0$, the numerical solution (\hat{p}_1, q_1) of (8.49) is exactly the same as the numerical solution of the symplectic Euler method (8.45) applied to the (unconstrained) Hamiltonian system (8.51). Therefore, Example 8.9 shows that the numerical solution (\hat{p}_1, q_1) is $\mathcal{O}(h^4)$ -close to the exact solution of (8.46), where in the definition of \tilde{H} the function H has to be replaced by \mathcal{H} of Eq. (8.50).

The projection step (8.8d,e) can be treated similarly. The solution μ of (8.8d), (8.9b) depends on \hat{p}_1, q_1 , and h (the dependence on p_0, q_0 can be omitted, because the relations (8.8a,b) allow us to express them in terms of \hat{p}_1, q_1 , and h). Due to the relation (8.8d) we can also consider μ as a function of p_1, q_1, h , i.e., $\mu = \mu(p_1, q_1, h)$. We now consider the Hamiltonian

$$\mathcal{G}(p, q) = g(q)^T \mu(p, q, h), \quad (8.52)$$

and the corresponding Hamiltonian system

$$\begin{aligned} q' &= g(q)^T \mu_p(p, q, h) \\ p' &= -G^T(q)\mu(p, q, h) - g(q)^T \mu_q(p, q, h). \end{aligned} \quad (8.53)$$

If $g(q_1) = 0$, the numerical approximation p_1 , computed from (8.8d), i.e., $p_1 = \hat{p}_1 - hG^T(q_1)\mu(p_1, q_1, h)$, is identical to the numerical solution of (8.45), applied to the system (8.53) with initial values (\hat{p}_1, q_1) . Again, we obtain from Example 8.9 that the numerical solution (\underline{p}_1, q_1) is $\mathcal{O}(h^4)$ -close to the exact solution of (8.46), where in the definition of \tilde{H} the function H has to be replaced by \mathcal{G} of Eq. (8.52). We summarize our findings in the following theorem.

Theorem 8.11. *Consider the one-step method (8.8) and assume that the initial values are consistent, i.e., $(p_0, q_0) \in \mathcal{M}$. Then it holds*

$$p_1 - \tilde{p}(h) = \mathcal{O}(h^4), \quad q_1 - \tilde{q}(h) = \mathcal{O}(h^4),$$

where $\tilde{p}(t), \tilde{q}(t)$ is the solution of the Hamiltonian system (8.46) with

$$\tilde{H} = \hat{H} + \hat{G} + \frac{h}{2} \{\hat{H}, \hat{G}\} + \frac{h^2}{12} \left(\{\hat{H}, \{\hat{H}, \hat{G}\}\} + \{\hat{G}, \{\hat{G}, \hat{H}\}\} \right)$$

where

$$\begin{aligned} \hat{H} &= \mathcal{H} - \frac{h}{2} \mathcal{H}_p \mathcal{H}_q + \frac{h^2}{12} \left(\mathcal{H}_{pp} \mathcal{H}_q^2 + \mathcal{H}_{qq} \mathcal{H}_p^2 + 4\mathcal{H}_{pq} \mathcal{H}_p \mathcal{H}_q \right) \\ \hat{G} &= \mathcal{G} - \frac{h}{2} \mathcal{G}_p \mathcal{G}_q + \frac{h^2}{12} \left(\mathcal{G}_{pp} \mathcal{G}_q^2 + \mathcal{G}_{qq} \mathcal{G}_p^2 + 4\mathcal{G}_{pq} \mathcal{G}_p \mathcal{G}_q \right), \end{aligned}$$

and \mathcal{H} and \mathcal{G} are given by (8.50) and (8.52), respectively. Here, the poisson bracket $\{H, G\}$ of two functions $H, G : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is given by $\{H, G\} := H_p G_q - H_q G_p$ (see Eq. (II.16.65)).

Proof. We consider the one-step method (8.8) as a composition of the mappings $(p_0, q_0) \mapsto (\hat{p}_1, q_1)$ and $(\hat{p}_1, q_1) \mapsto (p_1, q_1)$. Neglecting terms of size $\mathcal{O}(h^4)$, both mappings can be interpreted as the h -flow of Hamiltonian systems. The statement thus follows from the Campbell-Baker-Hausdorff Formula (II.16.83). \square

Backward Error Analysis for the Method (8.19). We consider the solution λ_0 of (8.19a,b), (8.9a) as a function of $p_{1/2}, q_0$, and h , i.e., $\lambda_0 = \lambda(p_{1/2}, q_0, h)$, and the solution μ_0 of (8.19d), (8.9b) as a function of p_1, q_1 and h , i.e., $\mu_0 = \mu(p_1, q_1, h)$. The method (8.19) can therefore be written as the composition of

$$\begin{aligned} p_{1/2} &= p_0 - \frac{h}{2} \left(H_q(p_{1/2}, q_0) + G^T(q_0) \lambda(p_{1/2}, q_0, h) \right) \\ q_1 &= q_0 + \frac{h}{2} \left(H_p(p_{1/2}, q_0) + H_p(p_{1/2}, q_1) \right) \\ \hat{p}_1 &= p_{1/2} - \frac{h}{2} \left(H_q(p_{1/2}, q_1) + G^T(q_1) \lambda(p_{1/2}, q_1, h) \right) \end{aligned} \quad (8.54)$$

with the projection step

$$p_1 = \hat{p}_1 - \frac{h}{2} G^T(q_1) \nu(p_1, q_1, h), \quad (8.55)$$

where $\nu(p_1, q_1, h) = \mu(p_1, q_1, h) - \lambda(p_{1/2}, q_1, h)$. We see that, for consistent initial values $(p_0, q_0) \in \mathcal{M}$, (8.54) is identical to (8.47) with $H(p, q)$ replaced by

$$\mathcal{H}(p, q) = H(p, q) + g(q)^T \lambda(p, q, h), \quad (8.56)$$

and the projection step (8.55) can be interpreted as method (8.45) with Hamiltonian function

$$\mathcal{G}(p, q) = \frac{1}{2} g(q)^T \nu(p, q, h). \quad (8.57)$$

In the same way as for the first order method we get:

Theorem 8.12. *Consider the method (8.19) and assume consistent initial values $(p_0, q_0) \in \mathcal{M}$. Then it holds*

$$p_1 - \tilde{p}(h) = \mathcal{O}(h^4), \quad q_1 - \tilde{q}(h) = \mathcal{O}(h^4),$$

where $\tilde{p}(t), \tilde{q}(t)$ is the solution of the Hamiltonian system (8.46) with

$$\tilde{H} = \hat{H} + \hat{G} + \frac{h}{2} \{ \hat{H}, \hat{G} \} + \frac{h^2}{12} \left(\{ \hat{H}, \{ \hat{H}, \hat{G} \} \} + \{ \hat{G}, \{ \hat{G}, \hat{H} \} \} \right)$$

where

$$\begin{aligned} \hat{H} &= \mathcal{H} + \frac{h^2}{24} \left(2\mathcal{H}_{qq}\mathcal{H}_p^2 - \mathcal{H}_{pp}\mathcal{H}_q^2 + 2\mathcal{H}_{pq}\mathcal{H}_p\mathcal{H}_q \right) \\ \hat{G} &= \mathcal{G} - \frac{h}{2}\mathcal{G}_p\mathcal{G}_q + \frac{h^2}{12} \left(\mathcal{G}_{pp}\mathcal{G}_q^2 + \mathcal{G}_{qq}\mathcal{G}_p^2 + 4\mathcal{G}_{pq}\mathcal{G}_p\mathcal{G}_q \right), \end{aligned}$$

and \mathcal{H} and \mathcal{G} are given by (8.56) and (8.57), respectively. \square

The above two theorems show that, for consistent initial values, the numerical solution of the considered methods is (up to a certain order) the exact solution of an unconstrained perturbed Hamiltonian system. The perturbed Hamiltonian is defined in a neighbourhood of the manifold, so that all backward error analysis results for ODEs can be applied.

Exercises

1. (Jay 1995). The system (1.46) is equivalent to

$$\begin{aligned} q' &= u \\ (M(q)u)' &= M_q(q)(u, u) + f(q, u) - G^T(q)\lambda \\ 0 &= g(q). \end{aligned} \tag{8.58}$$

In the case where (1.46) is obtained from the Lagrangian function $\mathcal{L}(q, \dot{q}) = \frac{1}{2}\dot{q}^T M(q)\dot{q} - U(q)$, show that $f(q, u)$ always contains the term $-M_q(q)(u, u)$ (Coriolis forces), which thus cancels out in the formulation (8.58).

2. Show that the example (2.1a-c) is of the form (8.1a-c) with Hamiltonian

$$H(p, q) = (p_1^2 + p_2^2)/2 + q_2.$$

If we compute λ from (2.3), and insert it into (2.1a,b), the resulting differential equation is no longer Hamiltonian.

3. Give a second proof of Theorem 8.1 by applying Theorem I.14.12.

Hint (Reich 1996). Let $\lambda = \lambda(p, q)$ be defined by (8.3b) and consider the unconstrained Hamiltonian system with Hamiltonian

$$H(p, q) + g(q)^T \lambda(p, q),$$

whose flow reduces to that of (8.1) along the constraint manifold \mathcal{M} .

4. Consider a partitioned Runge-Kutta method applied to a partitioned ordinary differential equation (without constraints). Suppose that both methods are based on the same quadrature formula of order p , that the first method satisfies $C(\eta), D(\xi)$, and that the second method satisfies $C(\hat{\eta}), D(\hat{\xi})$. Prove that the pair has order

$$\min\left(p, 2\min(\eta, \hat{\eta}) + 2, \min(\eta, \hat{\eta}) + \min(\xi, \hat{\xi}) + 2, \min(\eta + \xi, \hat{\eta} + \hat{\xi}) + 1\right).$$

Conclude that the Lobatto IIIA-III B pair has order $2s - 2$.

Hint. Apply the ideas of the proof of Theorem II.7.4 for the verification of the order conditions (Sect. II.15).

5. Compute a first integral of the differential equation (8.43). What is the reason for the existence of such an invariant?

Hint. With the transformation $u = e^p$, $v = e^q$ you will get a Hamiltonian system.

Result. $\tilde{I}(u, v) = I(u, v) + h((u + v)^2 - 10u - 8v + 8 \ln u + 2 \ln v)/4$.



iso geht alles zu Ende allhier:
Feder, Tinte, Tobak und auch wir.
Zum letztenmal wird eingetunkt,
Dann kommt der große

schwarze



(W. Busch, Bilder zur Jobsiade 1872)

Appendix. Fortran Codes

During the preparation of this book several programs have been developed for solving stiff and differential-algebraic problems of the form

$$My' = f(x, y), \quad y(x_0) = y_0, \quad (\text{A.1})$$

where M is a constant square matrix. If M is singular, the problem is differential-algebraic. In this case the initial values have to be consistent.

The implicit Runge-Kutta code RADAU5 and its extension RADAUP can be applied to higher index (≥ 2) problems as well, whereas the Rosenbrock code RODAS and the extrapolation code SEULEX are suited for explicit stiff differential equations and index 1 problems. The codes SDIRK4, ROS4, and SODEX are still available, but have not been updated.

In the case where M is not a constant matrix, suitable transformations and/or introduction of new variables allow us to bring every implicit differential equation to the form (A.1). If the problem is originally in one of the following forms

$$B(y)y' = f(x, y), \quad y'' = f(x, y, y'), \quad B(y)y'' = f(x, y, y'),$$

or the like, then the efficiency of the code can be increased by setting some parameters. This will be explained later in this appendix.

Communication with the code during integration can be done with help of the user-supplied subroutine SOLOUT. This is illustrated in the driver below. Further applications of this subroutine are discussed at the end of this appendix.

Experiences with all of our codes are welcome. The programs can be obtained by anonymous ftp (from “ftp.unige.ch” in the directory “pub/doc/math” or from “http://www.unige.ch/math/folks/haier/”).

Address: Section de Mathématiques, Case postale 240, CH-1211 Genève 24,
Switzerland

E-mail: Ernst.Hairer@math.unige.ch Gerhard.Wanner@math.unige.ch

Driver for the Code RADAU5

“The van der Pol equation problem is so much harder than the rest . . .”
(L.F. Shampine 1987)

We consider the van der Pol equation

$$\begin{aligned} y_1' &= y_2 & y_1(0) &= 2 \\ y_2' &= ((1 - y_1^2)y_2 - y_1)/\varepsilon & y_2(0) &= -0.66 \end{aligned}$$

with $\varepsilon = 10^{-6}$ on the interval $[0,2]$. The subroutines FVPOL, JVPOL compute the right-hand side of this differential equation and its Jacobian. The subroutine SOLOUT is used to print the solution at equidistant points.

```

C -----
C link driver radau5 decsol dc-decsol or
C link driver radau5 lapack lapackc dc-lapack
C -----
      IMPLICIT REAL*8 (A-H,O-Z)
C --- PARAMETERS FOR RADAU5 (FULL JACOBIAN)
      PARAMETER (ND=2,LWORK=4*ND*ND+12*ND+20,LIWORK=3*ND+20)
      DIMENSION Y(ND),WORK(LWORK),IWORK(LIWORK)
      EXTERNAL FVPOL,JVPOL,SOLOUT
C --- PARAMETER IN THE DIFFERENTIAL EQUATION
      RPAR=1.0D-6
C --- DIMENSION OF THE SYSTEM
      N=2
C --- COMPUTE THE JACOBIAN ANALYTICALLY
      IJAC=1
C --- JACOBIAN IS A FULL MATRIX
      MLJAC=N
C --- DIFFERENTIAL EQUATION IS IN EXPLICIT FORM
      IMAS=0
C --- OUTPUT ROUTINE IS USED DURING INTEGRATION
      IOUT=1
C --- INITIAL VALUES
      X=0.0D0
      Y(1)=2.0D0
      Y(2)=-0.66D0
C --- ENDPOINT OF INTEGRATION
      XEND=2.0D0
C --- REQUIRED TOLERANCE
      RTOL=1.0D-4
      ATOL=1.0D0*RTOL
      ITOL=0
C --- INITIAL STEP SIZE
      H=1.0D-6
C --- SET DEFAULT VALUES
      DO I=1,20
         IWORK(I)=0
         WORK(I)=0.0D0
      END DO
C --- CALL OF THE SUBROUTINE RADAU5
      CALL RADAU5(N,FVPOL,X,Y,XEND,H,
+              RTOL,ATOL,ITOL,
+              JVPOL,IJAC,MLJAC,MUJAC,
+              FVPOL,IMAS,MLMAS,MUMAS,
+              SOLOUT,IOUT,
+              WORK,LWORK,IWORK,LIWORK,RPAR,IPAR,IDID)

```

```

C --- PRINT FINAL SOLUTION
      WRITE (6,99) X,Y(1),Y(2)
99    FORMAT(1X,'X =',F5.2,'      Y =',2E18.10)
C --- PRINT STATISTICS
      WRITE (6,90) RTOL
90    FORMAT('      rtol=',D8.2)
      WRITE (6,91) (IWORK(J),J=14,20)
91    FORMAT(' fcn=',I5,' jac=',I4,' step=',I4,' accpt=',I4,
+          ' reject=',I3,' dec=',I4,' sol=',I5)
      STOP
      END

C
      SUBROUTINE SOLOUT (NR,XOLD,X,Y,CONT,LRC,N,RPAR,IPAR,IRTRN)
C --- PRINTS SOLUTION AT EQUIDISTANT OUTPUT-POINTS BY USING "CONTR5"
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION Y(N),CONT(LRC)
      COMMON /INTERN/XOUT
      IF (NR.EQ.1) THEN
        WRITE (6,99) X,Y(1),Y(2),NR-1
        XOUT=0.2D0
      ELSE
10     CONTINUE
        IF (X.GE.XOUT) THEN
C --- CONTINUOUS OUTPUT FOR RADAU5
          WRITE (6,99) XOUT,CONTR5(1,XOUT,CONT,LRC),
+              CONTR5(2,XOUT,CONT,LRC),NR-1
          XOUT=XOUT+0.2D0
          GOTO 10
        END IF
      END IF
99    FORMAT(1X,'X =',F5.2,'      Y =',2E18.10,'      NSTEP =',I4)
      RETURN
      END

C
      SUBROUTINE FVPOL(N,X,Y,F,RPAR,IPAR)
C --- RIGHT-HAND SIDE OF VAN DER POL'S EQUATION
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION Y(N),F(N)
      F(1)=Y(2)
      F(2)=((1-Y(1)**2)*Y(2)-Y(1))/RPAR
      RETURN
      END

C
      SUBROUTINE JVPOL(N,X,Y,DFY,LDFY,RPAR,IPAR)
C --- JACOBIAN OF VAN DER POL'S EQUATION
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION Y(N),DFY(LDFY,N)
      DFY(1,1)=0.0D0
      DFY(1,2)=1.0D0
      DFY(2,1)=(-2.0D0*Y(1)*Y(2)-1.0D0)/RPAR
      DFY(2,2)=(1.0D0-Y(1)**2)/RPAR
      RETURN
      END

```

The result, obtained on a Sun SPARKstation 20, is the following:

X = 0.00	Y = 0.2000000000E+01	-0.6600000000E+00	NSTEP = 0
X = 0.20	Y = 0.1858210825E+01	-0.7575052373E+00	NSTEP = 10
X = 0.40	Y = 0.1693217727E+01	-0.9068995621E+00	NSTEP = 11
X = 0.60	Y = 0.1484573110E+01	-0.1233017457E+01	NSTEP = 13
X = 0.80	Y = 0.1083921362E+01	-0.6195010714E+01	NSTEP = 21

```

X = 1.00   Y = -0.1863641256E+01   0.7535196392E+00   NSTEP = 144
X = 1.20   Y = -0.1699715970E+01   0.8997232240E+00   NSTEP = 145
X = 1.40   Y = -0.1493380698E+01   0.1213958018E+01   NSTEP = 147
X = 1.60   Y = -0.1120822309E+01   0.4373266499E+01   NSTEP = 153
X = 1.80   Y =  0.1869064482E+01   -0.7496053261E+00   NSTEP = 275
X = 2.00   Y =  0.1706171005E+01   -0.8928020961E+00   NSTEP = 276
X = 2.00   Y =  0.1706171005E+01   -0.8928020961E+00
      rtol=0.10D-03
fcn= 2263 jac= 182 step= 293 accpt= 276 rejct=  9 dec= 251 sol= 662

```

Subroutine RADAU5

Implicit Runge-Kutta code based on the 3-stage Radau IIA method, given in Table IV.5.6. Details on the implementation are described in Section IV.8.

```

      SUBROUTINE RADAU5(N,FCN,X,Y,XEND,H,
+          RTOL,ATOL,ITOL,
+          JAC,IJAC,MLJAC,MUJAC,
+          MAS,IMAS,MLMAS,MUMAS,
+          SOLOUT,IOUT,
+          WORK,LWORK,IWORK,LIWORK,RPAR,IPAR,IDID)
C -----
C      NUMERICAL SOLUTION OF A STIFF (OR DIFFERENTIAL ALGEBRAIC)
C      SYSTEM OF FIRST ORDER ORDINARY DIFFERENTIAL EQUATIONS
C          M*Y'=F(X,Y).
C      THE SYSTEM CAN BE (LINEARLY) IMPLICIT (MASS-MATRIX M .NE. I)
C      OR EXPLICIT (M=I).
C      THE METHOD USED IS AN IMPLICIT RUNGE-KUTTA METHOD (RADAU IIA)
C      OF ORDER 5 WITH STEP SIZE CONTROL AND CONTINUOUS OUTPUT.
C      C.F. SECTION IV.8
C
C      AUTHORS: E. HAIRER AND G. WANNER
C              UNIVERSITE DE GENEVE, DEPT. DE MATHEMATIQUES
C              CH-1211 GENEVE 24, SWITZERLAND
C              E-MAIL: HAIRER@DIVSUN.UNIGE.CH, WANNER@DIVSUN.UNIGE.CH
C
C      THIS CODE IS PART OF THE BOOK:
C          E. HAIRER AND G. WANNER, SOLVING ORDINARY DIFFERENTIAL
C          EQUATIONS II. STIFF AND DIFFERENTIAL-ALGEBRAIC PROBLEMS.
C          SPRINGER SERIES IN COMPUTATIONAL MATHEMATICS 14,
C          SPRINGER-VERLAG 1991, SECOND EDITION 1996.
C
C      VERSION OF SEPTEMBER 30, 1995
C
C      INPUT PARAMETERS
C      -----
C      N          DIMENSION OF THE SYSTEM
C
C      FCN        NAME (EXTERNAL) OF SUBROUTINE COMPUTING THE
C                VALUE OF F(X,Y):
C                SUBROUTINE FCN(N,X,Y,F,RPAR,IPAR)
C                REAL*8 X,Y(N),F(N)
C                F(1)=...   ETC.
C                RPAR, IPAR (SEE BELOW)
C
C      X          INITIAL X-VALUE
C
C      Y(N)       INITIAL VALUES FOR Y

```

```

C
C  XEND      FINAL X-VALUE (XEND-X MAY BE POSITIVE OR NEGATIVE)
C
C  H         INITIAL STEP SIZE GUESS;
C           FOR STIFF EQUATIONS WITH INITIAL TRANSIENT,
C           H=1.DO/(NORM OF F'), USUALLY 1.D-3 OR 1.D-5, IS GOOD.
C           THIS CHOICE IS NOT VERY IMPORTANT, THE STEP SIZE IS
C           QUICKLY ADAPTED. (IF H=0.DO, THE CODE PUTS H=1.D-6).
C
C  RTOL,ATOL  RELATIVE AND ABSOLUTE ERROR TOLERANCES. THEY
C           CAN BE BOTH SCALARS OR ELSE BOTH VECTORS OF LENGTH N.
C
C  ITOL      SWITCH FOR RTOL AND ATOL:
C           ITOL=0:  BOTH RTOL AND ATOL ARE SCALARS.
C                   THE CODE KEEPS, ROUGHLY, THE LOCAL ERROR OF
C                   Y(I) BELOW RTOL*ABS(Y(I))+ATOL
C           ITOL=1:  BOTH RTOL AND ATOL ARE VECTORS.
C                   THE CODE KEEPS THE LOCAL ERROR OF Y(I) BELOW
C                   RTOL(I)*ABS(Y(I))+ATOL(I).
C
C  JAC       NAME (EXTERNAL) OF THE SUBROUTINE WHICH COMPUTES
C           THE PARTIAL DERIVATIVES OF F(X,Y) WITH RESPECT TO Y
C           (THIS ROUTINE IS ONLY CALLED IF IJAC=1; SUPPLY
C           A DUMMY SUBROUTINE IN THE CASE IJAC=0).
C           FOR IJAC=1, THIS SUBROUTINE MUST HAVE THE FORM
C           SUBROUTINE JAC(N,X,Y,DFY,LDFY,RPAR,IPAR)
C           REAL*8 X,Y(N),DFY(LDFY,N)
C           DFY(1,1)= ...
C           LDFY, THE COLUMN-LENGTH OF THE ARRAY, IS
C           FURNISHED BY THE CALLING PROGRAM.
C           IF (MLJAC.EQ.N) THE JACOBIAN IS SUPPOSED TO
C           BE FULL AND THE PARTIAL DERIVATIVES ARE
C           STORED IN DFY AS
C               DFY(I,J) = PARTIAL F(I) / PARTIAL Y(J)
C           ELSE, THE JACOBIAN IS TAKEN AS BANDED AND
C           THE PARTIAL DERIVATIVES ARE STORED
C           DIAGONAL-WISE AS
C               DFY(I-J+MUJAC+1,J) = PARTIAL F(I) / PARTIAL Y(J).
C
C  IJAC      SWITCH FOR THE COMPUTATION OF THE JACOBIAN:
C           IJAC=0:  JACOBIAN IS COMPUTED INTERNALLY BY FINITE
C                   DIFFERENCES, SUBROUTINE "JAC" IS NEVER CALLED.
C           IJAC=1:  JACOBIAN IS SUPPLIED BY SUBROUTINE JAC.
C
C  MLJAC     SWITCH FOR THE BANDED STRUCTURE OF THE JACOBIAN:
C           MLJAC=N: JACOBIAN IS A FULL MATRIX. THE LINEAR
C                   ALGEBRA IS DONE BY FULL-MATRIX GAUSS-ELIMINATION.
C           0<=MLJAC<N: MLJAC IS THE LOWER BANDWIDTH OF JACOBIAN
C                   MATRIX (>= NUMBER OF NON-ZERO DIAGONALS BELOW
C                   THE MAIN DIAGONAL).
C
C  MUJAC     UPPER BANDWIDTH OF JACOBIAN MATRIX (>= NUMBER OF NON-
C           ZERO DIAGONALS ABOVE THE MAIN DIAGONAL).
C           NEED NOT BE DEFINED IF MLJAC=N.
C
C  ----  MAS,IMAS,MLMAS, AND MUMAS HAVE ANALOG MEANINGS  ----
C  ----  FOR THE "MASS MATRIX" (THE MATRIX "M" OF SECTION IV.8):  -
C
C  MAS      NAME (EXTERNAL) OF SUBROUTINE COMPUTING THE MASS-
C           MATRIX M.
C           IF IMAS=0, THIS MATRIX IS ASSUMED TO BE THE IDENTITY
C           MATRIX AND NEEDS NOT TO BE DEFINED;

```

```

C          SUPPLY A DUMMY SUBROUTINE IN THIS CASE.
C          IF IMAS=1, THE SUBROUTINE MAS IS OF THE FORM
C          SUBROUTINE MAS(N,AM,LMAS,RPAR,IPAR)
C          REAL*8 AM(LMAS,N)
C          AM(1,1)= ....
C          IF (MLMAS.EQ.N) THE MASS-MATRIX IS STORED
C          AS FULL MATRIX LIKE
C          AM(I,J) = M(I,J)
C          ELSE, THE MATRIX IS TAKEN AS BANDED AND STORED
C          DIAGONAL-WISE AS
C          AM(I-J+MUMAS+1,J) = M(I,J).
C
C  IMAS      GIVES INFORMATION ON THE MASS-MATRIX:
C            IMAS=0:  M IS SUPPOSED TO BE THE IDENTITY
C            MATRIX, MAS IS NEVER CALLED.
C            IMAS=1:  MASS-MATRIX IS SUPPLIED.
C
C  MLMAS     SWITCH FOR THE BANDED STRUCTURE OF THE MASS-MATRIX:
C            MLMAS=N: THE FULL MATRIX CASE. THE LINEAR
C            ALGEBRA IS DONE BY FULL-MATRIX GAUSS-ELIMINATION.
C            0<=MLMAS<N: MLMAS IS THE LOWER BANDWIDTH OF THE
C            MATRIX (>= NUMBER OF NON-ZERO DIAGONALS BELOW
C            THE MAIN DIAGONAL).
C            MLMAS IS SUPPOSED TO BE .LE. MLJAC.
C
C  MUMAS     UPPER BANDWIDTH OF MASS-MATRIX (>= NUMBER OF NON-
C            ZERO DIAGONALS ABOVE THE MAIN DIAGONAL).
C            NEED NOT BE DEFINED IF MLMAS=N.
C            MUMAS IS SUPPOSED TO BE .LE. MUJAC.
C
C  SOLOUT    NAME (EXTERNAL) OF SUBROUTINE PROVIDING THE
C            NUMERICAL SOLUTION DURING INTEGRATION.
C            IF IOUT=1, IT IS CALLED AFTER EVERY SUCCESSFUL STEP.
C            SUPPLY A DUMMY SUBROUTINE IF IOUT=0.
C            IT MUST HAVE THE FORM
C            SUBROUTINE SOLOUT (NR,XOLD,X,Y,CONT,LRC,N,
C                               RPAR,IPAR,IRTRN)
C            REAL*8 X,Y(N),CONT(LRC)
C            ....
C            SOLOUT FURNISHES THE SOLUTION "Y" AT THE NR-TH
C            GRID-POINT "X" (THEREBY THE INITIAL VALUE IS
C            THE FIRST GRID-POINT).
C            "XOLD" IS THE PRECEEDING GRID-POINT.
C            "IRTRN" SERVES TO INTERRUPT THE INTEGRATION. IF IRTRN
C            IS SET <0, RADAU5 RETURNS TO THE CALLING PROGRAM.
C
C          ----- CONTINUOUS OUTPUT: -----
C          DURING CALLS TO "SOLOUT", A CONTINUOUS SOLUTION
C          FOR THE INTERVAL [XOLD,X] IS AVAILABLE THROUGH
C          THE FUNCTION
C          >>> CONTR5(I,S,CONT,LRC) <<<
C          WHICH PROVIDES AN APPROXIMATION TO THE I-TH
C          COMPONENT OF THE SOLUTION AT THE POINT S. THE VALUE
C          S SHOULD LIE IN THE INTERVAL [XOLD,X].
C          DO NOT CHANGE THE ENTRIES OF CONT(LRC), IF THE
C          DENSE OUTPUT FUNCTION IS USED.
C
C  IOUT      SWITCH FOR CALLING THE SUBROUTINE SOLOUT:
C            IOUT=0:  SUBROUTINE IS NEVER CALLED
C            IOUT=1:  SUBROUTINE IS AVAILABLE FOR OUTPUT.
C
C  WORK      ARRAY OF WORKING SPACE OF LENGTH "LWORK".

```

```

C      WORK(1), WORK(2),..., WORK(20) SERVE AS PARAMETERS
C      FOR THE CODE. FOR STANDARD USE OF THE CODE
C      WORK(1),...,WORK(20) MUST BE SET TO ZERO BEFORE
C      CALLING. SEE BELOW FOR A MORE SOPHISTICATED USE.
C      WORK(21),...,WORK(LWORK) SERVE AS WORKING SPACE
C      FOR ALL VECTORS AND MATRICES.
C      "LWORK" MUST BE AT LEAST
C          N*(LJAC+LMAS+3*LE+12)+20
C      WHERE
C          LJAC=N          IF MLJAC=N (FULL JACOBIAN)
C          LJAC=MLJAC+MUJAC+1 IF MLJAC<N (BANDED JAC.)
C      AND
C          LMAS=0          IF IMAS=0
C          LMAS=N          IF IMAS=1 AND MLMAS=N (FULL)
C          LMAS=MLMAS+MUMAS+1 IF MLMAS<N (BANDED MASS-M.)
C      AND
C          LE=N            IF MLJAC=N (FULL JACOBIAN)
C          LE=2*MLJAC+MUJAC+1 IF MLJAC<N (BANDED JAC.)
C
C      IN THE USUAL CASE WHERE THE JACOBIAN IS FULL AND THE
C      MASS-MATRIX IS THE INDENTITY (IMAS=0), THE MINIMUM
C      STORAGE REQUIREMENT IS
C          LWORK = 4*N*N+12*N+20.
C      IF IWORK(9)=M1>0 THEN "LWORK" MUST BE AT LEAST
C          N*(LJAC+12)+(N-M1)*(LMAS+3*LE)+20
C      WHERE IN THE DEFINITIONS OF LJAC, LMAS AND LE THE
C      NUMBER N CAN BE REPLACED BY N-M1.
C
C      LWORK      DECLARED LENGHT OF ARRAY "WORK".
C
C      IWORK      INTEGER WORKING SPACE OF LENGHT "LIWORK".
C      IWORK(1),IWORK(2),...,IWORK(20) SERVE AS PARAMETERS
C      FOR THE CODE. FOR STANDARD USE, SET IWORK(1),...,
C      IWORK(20) TO ZERO BEFORE CALLING.
C      IWORK(21),...,IWORK(LIWORK) SERVE AS WORKING AREA.
C      "LIWORK" MUST BE AT LEAST 3*N+20.
C
C      LIWORK     DECLARED LENGHT OF ARRAY "IWORK".
C
C      RPAR, IPAR  REAL AND INTEGER PARAMETERS (OR PARAMETER ARRAYS) WHICH
C      CAN BE USED FOR COMMUNICATION BETWEEN YOUR CALLING
C      PROGRAM AND THE FCN, JAC, MAS, SOLOUT SUBROUTINES.
C
C      -----
C
C      SOPHISTICATED SETTING OF PARAMETERS
C      -----
C
C      SEVERAL PARAMETERS OF THE CODE ARE TUNED TO MAKE IT WORK
C      WELL. THEY MAY BE DEFINED BY SETTING WORK(1),...
C      AS WELL AS IWORK(1),... DIFFERENT FROM ZERO.
C      FOR ZERO INPUT, THE CODE CHOOSES DEFAULT VALUES:
C
C      IWORK(1)    IF IWORK(1).NE.0, THE CODE TRANSFORMS THE JACOBIAN
C      MATRIX TO HESSENBERG FORM. THIS IS PARTICULARLY
C      ADVANTAGEOUS FOR LARGE SYSTEMS WITH FULL JACOBIAN.
C      IT DOES NOT WORK FOR BANDED JACOBIAN (MLJAC<N)
C      AND NOT FOR IMPLICIT SYSTEMS (IMAS=1).
C
C      IWORK(2)    THIS IS THE MAXIMAL NUMBER OF ALLOWED STEPS.
C      THE DEFAULT VALUE (FOR IWORK(2)=0) IS 100000.
C
C      IWORK(3)    THE MAXIMUM NUMBER OF NEWTON ITERATIONS FOR THE

```



```

C          SOLUTION OF THE IMPLICIT SYSTEM IN EACH STEP.
C          THE DEFAULT VALUE (FOR IWORK(3)=0) IS 7.
C
C          IWORK(4) IF IWORK(4).EQ.0 THE EXTRAPOLATED COLLOCATION SOLUTION
C          IS TAKEN AS STARTING VALUE FOR NEWTON'S METHOD.
C          IF IWORK(4).NE.0 ZERO STARTING VALUES ARE USED.
C          THE LATTER IS RECOMMENDED IF NEWTON'S METHOD HAS
C          DIFFICULTIES WITH CONVERGENCE (THIS IS THE CASE WHEN
C          NSTEP IS LARGER THAN NACCPT + NREJCT; SEE OUTPUT PARAM.).
C          DEFAULT IS IWORK(4)=0.
C
C          THE FOLLOWING 3 PARAMETERS ARE IMPORTANT FOR
C          DIFFERENTIAL-ALGEBRAIC SYSTEMS OF INDEX > 1.
C          THE FUNCTION-SUBROUTINE SHOULD BE WRITTEN SUCH THAT
C          THE INDEX 1,2,3 VARIABLES APPEAR IN THIS ORDER.
C          IN ESTIMATING THE ERROR THE INDEX 2 VARIABLES ARE
C          MULTIPLIED BY H, THE INDEX 3 VARIABLES BY H**2.
C
C          IWORK(5) DIMENSION OF THE INDEX 1 VARIABLES (MUST BE > 0). FOR
C          ODE'S THIS EQUALS THE DIMENSION OF THE SYSTEM.
C          DEFAULT IWORK(5)=N.
C
C          IWORK(6) DIMENSION OF THE INDEX 2 VARIABLES. DEFAULT IWORK(6)=0.
C
C          IWORK(7) DIMENSION OF THE INDEX 3 VARIABLES. DEFAULT IWORK(7)=0.
C
C          IWORK(8) SWITCH FOR STEP SIZE STRATEGY
C          IF IWORK(8).EQ.1 MOD. PREDICTIVE CONTROLLER (GUSTAFSSON)
C          IF IWORK(8).EQ.2 CLASSICAL STEP SIZE CONTROL
C          THE DEFAULT VALUE (FOR IWORK(8)=0) IS IWORK(8)=1.
C          THE CHOICE IWORK(8).EQ.1 SEEMS TO PRODUCE SAFER RESULTS;
C          FOR SIMPLE PROBLEMS, THE CHOICE IWORK(8).EQ.2 PRODUCES
C          OFTEN SLIGHTLY FASTER RUNS
C
C          IF THE DIFFERENTIAL SYSTEM HAS THE SPECIAL STRUCTURE THAT
C           $Y(I)' = Y(I+M2)$  FOR  $I=1, \dots, M1$ ,
C          WITH  $M1$  A MULTIPLE OF  $M2$ , A SUBSTANTIAL GAIN IN COMPUTERTIME
C          CAN BE ACHIEVED BY SETTING THE PARAMETERS IWORK(9) AND IWORK(10).
C          E.G., FOR SECOND ORDER SYSTEMS  $P'=V$ ,  $V'=G(P,V)$ , WHERE  $P$  AND  $V$  ARE
C          VECTORS OF DIMENSION  $N/2$ , ONE HAS TO PUT  $M1=M2=N/2$ .
C          FOR  $M1>0$  SOME OF THE INPUT PARAMETERS HAVE DIFFERENT MEANINGS:
C          - JAC: ONLY THE ELEMENTS OF THE NON-TRIVIAL PART OF THE
C          JACOBIAN HAVE TO BE STORED
C          IF (MLJAC.EQ.N-M1) THE JACOBIAN IS SUPPOSED TO BE FULL
C           $DFY(I,J) = \text{PARTIAL } F(I+M1) / \text{PARTIAL } Y(J)$ 
C          FOR  $I=1, N-M1$  AND  $J=1, N$ .
C          ELSE, THE JACOBIAN IS BANDED ( $M1 = M2 * MM$ )
C           $DFY(I-J+MUJAC+1, J+K*M2) = \text{PARTIAL } F(I+M1) / \text{PARTIAL } Y(J+K*M2)$ 
C          FOR  $I=1, MLJAC+MUJAC+1$  AND  $J=1, M2$  AND  $K=0, MM$ .
C          - MLJAC:  $MLJAC=N-M1$ : IF THE NON-TRIVIAL PART OF THE JACOBIAN IS FULL
C           $0 < MLJAC < N-M1$ : IF THE  $(MM+1)$  SUBMATRICES (FOR  $K=0, MM$ )
C           $\text{PARTIAL } F(I+M1) / \text{PARTIAL } Y(J+K*M2)$ ,  $I, J=1, M2$ 
C          ARE BANDED, MLJAC IS THE MAXIMAL LOWER BANDWIDTH
C          OF THESE  $MM+1$  SUBMATRICES
C          - MUJAC: MAXIMAL UPPER BANDWIDTH OF THESE  $MM+1$  SUBMATRICES
C          NEED NOT BE DEFINED IF  $MLJAC=N-M1$ 
C          - MAS: IF IMAS=0 THIS MATRIX IS ASSUMED TO BE THE IDENTITY AND
C          NEED NOT BE DEFINED. SUPPLY A DUMMY SUBROUTINE IN THIS CASE.
C          IT IS ASSUMED THAT ONLY THE ELEMENTS OF RIGHT LOWER BLOCK OF
C          DIMENSION  $N-M1$  DIFFER FROM THAT OF THE IDENTITY MATRIX.
C          IF (MLMAS.EQ.N-M1) THIS SUBMATRIX IS SUPPOSED TO BE FULL
C           $AM(I,J) = M(I+M1, J+M1)$  FOR  $I=1, N-M1$  AND  $J=1, N-M1$ .

```

```

C      ELSE, THE MASS MATRIX IS BANDED
C      AM(I-J+MUMAS+1,J) = M(I+M1,J+M1)
C      - MLMAS: MLMAS=N-M1: IF THE NON-TRIVIAL PART OF M IS FULL
C      0<=MLMAS<N-M1: LOWER BANDWIDTH OF THE MASS MATRIX
C      - MUMAS: UPPER BANDWIDTH OF THE MASS MATRIX
C      NEED NOT BE DEFINED IF MLMAS=N-M1
C
C      IWORK(9)  THE VALUE OF M1.      DEFAULT M1=0.
C
C      IWORK(10) THE VALUE OF M2.      DEFAULT M2=M1.
C
C      -----
C
C      WORK(1)   UROUND, THE ROUNDING UNIT, DEFAULT 1.D-16.
C
C      WORK(2)   THE SAFETY FACTOR IN STEP SIZE PREDICTION,
C      DEFAULT 0.9D0.
C
C      WORK(3)   DECIDES WHETHER THE JACOBIAN SHOULD BE RECOMPUTED;
C      INCREASE WORK(3), TO 0.1 SAY, WHEN JACOBIAN EVALUATIONS
C      ARE COSTLY. FOR SMALL SYSTEMS WORK(3) SHOULD BE SMALLER
C      (0.001D0, SAY). NEGATIV WORK(3) FORCES THE CODE TO
C      COMPUTE THE JACOBIAN AFTER EVERY ACCEPTED STEP.
C      DEFAULT 0.001D0.
C
C      WORK(4)   STOPPING CRITERION FOR NEWTON'S METHOD, USUALLY CHOSEN <1.
C      SMALLER VALUES OF WORK(4) MAKE THE CODE SLOWER, BUT SAFER.
C      DEFAULT 0.03D0.
C
C      WORK(5) AND WORK(6) : IF WORK(5) < HNEW/HOLD < WORK(6), THEN THE
C      STEP SIZE IS NOT CHANGED. THIS SAVES, TOGETHER WITH A
C      LARGE WORK(3), LU-DECOMPOSITIONS AND COMPUTING TIME FOR
C      LARGE SYSTEMS. FOR SMALL SYSTEMS ONE MAY HAVE
C      WORK(5)=1.D0, WORK(6)=1.2D0, FOR LARGE FULL SYSTEMS
C      WORK(5)=0.99D0, WORK(6)=2.D0 MIGHT BE GOOD.
C      DEFAULTS WORK(5)=1.D0, WORK(6)=1.2D0 .
C
C      WORK(7)   MAXIMAL STEP SIZE, DEFAULT XEND-X.
C
C      WORK(8), WORK(9)  PARAMETERS FOR STEP SIZE SELECTION
C      THE NEW STEP SIZE IS CHOSEN SUBJECT TO THE RESTRICTION
C      WORK(8) <= HNEW/HOLD <= WORK(9)
C      DEFAULT VALUES: WORK(8)=0.2D0, WORK(9)=8.D0
C
C      -----
C
C      OUTPUT PARAMETERS
C      -----
C
C      X          X-VALUE FOR WHICH THE SOLUTION HAS BEEN COMPUTED
C      (AFTER SUCCESSFUL RETURN X=XEND).
C
C      Y(N)       NUMERICAL SOLUTION AT X
C
C      H          PREDICTED STEP SIZE OF THE LAST ACCEPTED STEP
C
C      IDID       REPORTS ON SUCCESSFULNESS UPON RETURN:
C      IDID= 1    COMPUTATION SUCCESSFUL,
C      IDID= 2    COMPUT. SUCCESSFUL (INTERRUPTED BY SOLOUT)
C      IDID=-1    INPUT IS NOT CONSISTENT,
C      IDID=-2    LARGER NMAX IS NEEDED,
C      IDID=-3    STEP SIZE BECOMES TOO SMALL,
C      IDID=-4    MATRIX IS REPEATEDLY SINGULAR.

```

```

C
C   IWORK(14)  NFCN      NUMBER OF FUNCTION EVALUATIONS (THOSE FOR NUMERICAL
C                   EVALUATION OF THE JACOBIAN ARE NOT COUNTED)
C   IWORK(15)  NJAC      NUMBER OF JACOBIAN EVALUATIONS (EITHER ANALYTICALLY
C                   OR NUMERICALLY)
C   IWORK(16)  NSTEP     NUMBER OF COMPUTED STEPS
C   IWORK(17)  NACCPT     NUMBER OF ACCEPTED STEPS
C   IWORK(18)  NREJCT     NUMBER OF REJECTED STEPS (DUE TO ERROR TEST),
C                   (STEP REJECTIONS IN THE FIRST STEP ARE NOT COUNTED)
C   IWORK(19)  NDEC       NUMBER OF LU-DECOMPOSITIONS OF BOTH MATRICES
C   IWORK(20)  NSOL       NUMBER OF FORWARD-BACKWARD SUBSTITUTIONS, OF BOTH
C                   SYSTEMS; THE NSTEP FORWARD-BACKWARD SUBSTITUTIONS,
C                   NEEDED FOR STEP SIZE SELECTION, ARE NOT COUNTED
C-----

```

Subroutine RADAUP

With the option `IWORK(11) = 3` this code is mathematically equivalent to RADAU5. The only difference is that explicit sums have been replaced by loops, and that the coefficients of the method have been put into arrays. This makes the code a little bit slower (in particular for small problems), but has the advantage that the coefficients of the method can be easily changed. At the moment, the coefficients of the Radau IIA methods of orders 5, 9, and 13 are available by setting `IWORK(11)` equal to 3, 5, and 7, respectively. The calling list is the same as for RADAU5.

```

SUBROUTINE RADAUP(N,FCN,X,Y,XEND,H,
+               RTOL,ATOL,ITOL,
+               JAC ,IJAC,MLJAC,MUJAC,
+               MAS ,IMAS,MLMAS,MUMAS,
+               SOLOUT,IOUT,
+               WORK,LWORK,IWORK,LIWORK,RPAR,IPAR,IDID)

```

Subroutine RODAS

This is an implementation of the Rosenbrock method described in Section VI.3. It also satisfies the algebraic order conditions and can thus be applied to differential-algebraic problems of index 1. The calling list is:

```

SUBROUTINE RODAS(N,FCN,IFCN,X,Y,XEND,H,
+               RTOL,ATOL,ITOL,
+               JAC ,IJAC,MLJAC,MUJAC,DFX,IDFX,
+               MAS ,IMAS,MLMAS,MUMAS,
+               SOLOUT,IOUT,
+               WORK,LWORK,IWORK,LIWORK,RPAR,IPAR,IDID)

```

Compared to RADAU5 we have three additional parameters. `IFCN` indicates whether the right-hand side $f(x,y)$ of the problem (A.1) is independent of x or not. In the case that f depends on x , the code needs the partial derivative $\partial f / \partial x$. This can be provided numerically (set `IDFX = 0` and supply a dummy subroutine for `DFX`) or analytically. In the latter case, one has to set `IDFX = 1` and one has to supply a

subroutine computing $\partial f / \partial x$. Of course, the meaning of the `WORK` and `IWORK` parameters are not all the same as for `RADAU5`. They are described in the comments of the code.

Subroutine SEULEX

This is an extrapolation code based on the linearly implicit Euler method (Sections IV.9 and VI.4). A dense output has been included in cooperation with A. Ostermann. The meaning of the input parameters is the same as for `RODAS`. The `WORK` and `IWORK` parameters are described in the comments of the code.

```

SUBROUTINE SEULEX(N,FCN,IFCN,X,Y,XEND,H,
+               RTOL,ATOL,ITOL,
+               JAC ,IJAC,MLJAC,MUJAC,
+               MAS ,IMAS,MLMAS,MUMAS,
+               SOLOUT,IOUT,
+               WORK,LWORK,IWORK,LIWORK,RPAR,IPAR,IDID)

```

Problems with Special Structure

If the first m_1 equations of (A.1) are of the form

$$y'_i = y_{i+m_2} \quad \text{for } i = 1, \dots, m_1 \quad (\text{A.2})$$

with m_1 being an integer multiple of m_2 , and the remaining equations do not depend explicitly on y'_{m_1+1}, \dots, y'_n , it is recommended to set the parameters `IWORK(9)` and `IWORK(10)` equal to m_1 and m_2 , respectively. This implies a more efficient treatment of the arising linear systems and is, in particular, advantageous for a large value of m_1 .

If `IWORK(9)` is set to a nonzero value, care has to be taken with the definition of the subroutines `JAC` and `MAS`. Only the nontrivial part of the Jacobian (i.e., the rows with indices $m_1 + 1, \dots, n$) have to be computed and stored in an array of dimension $(n - m_1) \times n$. Similarly, only the right lower block (of dimension $n - m_1$) of the matrix M has to be defined in the subroutine `MAS`. However, the subroutine `FCN` must contain the definition of all components of $f(x, y)$, in particular also the statement `F(I) = Y(I+M2)` for $I=1, \dots, M1$. Banded options are still possible. Typical situations, where (A.2) arises, are the following:

$y'' = f(x, y, y')$. With the new variable $z = y'$ the system becomes

$$\begin{aligned} y' &= z \\ z' &= f(x, y, z), \end{aligned}$$

which is of the form (A.1). If $y \in \mathbb{R}^m$, both parameters `IWORK(9)` and `IWORK(10)` have to be set equal to m . Banded option can be used, if both $\partial f / \partial y$ and $\partial f / \partial y'$ are banded.

$C(x, y)y' = f(x, y)$. Again we introduce $z = y'$, so that this problem becomes equivalent to

$$\begin{aligned}y' &= z \\ 0 &= C(x, y)z - f(x, y).\end{aligned}$$

Both parameters `IWORK(9)` and `IWORK(10)` have to be set equal to the dimension of y . If only a few components of y' are multiplied by non-constant terms, then it may be more efficient to introduce new variables only for these components.

$C(x, y)y'' = f(x, y, y')$. With the new variables $z = y'$ and $u = z' = y''$, this problem can be written in the form (A.1) as follows

$$\begin{aligned}y' &= z \\ z' &= u \\ 0 &= C(x, y)u - f(x, y, z).\end{aligned}$$

Here m_2 is equal to the dimension of y , and $m_1 = 2m_2$.

Use of SOLOUT and of Dense Output

The subroutine SOLOUT, supplied by the user, is called after every accepted step and provides the solution over the whole step (dense output). This possibility can be used for tabulating the solution at prescribed output points (see the driver for RADAU5 above) or for graphical presentation of the solution. Further applications are the following:

Event location. Suppose we want to determine x such that $g(x, y(x)) = 0$, where $y(x)$ is the solution of (A.1). During integration one can check in the subroutine SOLOUT whether the values $g(x_{i-1}, y_{i-1})$ and $g(x_i, y_i)$ change sign. If this occurs, the dense output (which is available for all of our codes) can be used to localize the zero of $g(x, y(x))$. This procedure is very useful for problems with discontinuous right-hand side (see Sect. II.6).

Projection. An efficient way for solving higher index differential-algebraic equations is index-reduction combined with projection. If one applies a stiff (or non-stiff) code straightforwardly to an index-reduced problem, the obtained numerical solution will suffer from the so-called “drift-off” effect. In order to avoid this drift-off, it is recommended to project the numerical solution after every step onto the solution manifold of the problem. This can be conveniently done with help of the subroutine SOLOUT.

Bibliography

This bibliography includes the publications referred to in the text. Italic numbers in square brackets following a reference indicate the sections where the reference is cited.

- R. Abraham, J.E. Marsden & T. Ratiu (1983): *Manifolds, Tensor Analysis, and Applications*. Applied Mathematical Sciences vol. 75, Springer-Verlag 1983; second edition 1988, 654 pp. [VII.1]
- M. Abramowitz & I.A. Stegun (1964): *Handbook of mathematical functions*. Dover, 1000 pages. [IV.2], [IV.4], [IV.12], [IV.13], [IV.14]
- C.A. Addison (1979): *Implementing a stiff method based upon the second derivative formulas*. Techn. Rep. 130/79, Dept. of Comput. Sc., Univ. of Toronto, Canada. [V.3], [V.5]
- R.C. Aiken ed. (1985): *Stiff computation*. Oxford, Univ. Press, 462pp. [IV.1], [IV.3], [IV.8], [V.5]
- G. Akilov, see L. Kantorovich & G. Akilov.
- R. Alexander (1977): *Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s*. SIAM J. Numer. Anal., vol. 14, pp. 1006-1021. [IV.3], [IV.6]
- R. Alexander (1997): *Reliability of software for stiff initial value problems*. To appear in SIAM J. Sci. Comput. [IV.10]
- T. Alishenas (1992): *Zur numerischen Behandlung, Stabilisierung durch Projektion und Modellierung mechanischer Systeme mit Nebenbedingungen und Invarianten*. Dissertation, Stockholm, TRITA-NA-9202. [VII.2]
- T. Alishenas & Ö. Ólafsson (1994): *Modeling and velocity stabilization of constrained mechanical systems*. BIT, vol. 34, pp. 455-483. [VII.2]
- R. Alt (1971): *Méthodes A-stables pour l'intégration de systèmes différentiels mal conditionnés*. Thèse, Univ. Paris VI. [IV.6]
- H.C. Andersen (1983): *Rattle: a "velocity" version of the Shake algorithm for molecular dynamics calculations*. J. Comput. Phys., vol. 52, pp. 24-34. [VII.8]
- G.C. Andrews, see also M.K. Ormrod & G.C. Andrews.
- C. Arévalo, C. Führer & G. Söderlind (1996): *Stabilized multistep methods for index 2 Euler-Lagrange DAEs*. BIT, vol. 36, pp. 1-13. [VII.6]
- S. Arimoto, see J. Nagumo, S. Arimoto & S. Yoshizawa.
- M. Arnold (1993): *Stability of numerical methods for differential-algebraic equations of higher index*. Applied Numerical Mathematics, vol. 13, pp. 5-14. [VII.1]

- M. Arnold (1995): *Half-explicit Runge-Kutta methods with explicit stages for differential-algebraic systems of index 2*. Submitted for publication. [VII.6], [VII.7]
- V.I. Arnol'd (1979): *Matematičeskie metody klassičeskoj mehaniki*. Nauka, Moskva; English translation: Springer Verlag 1984, 1989. [VII.1]
- W.E. Arnoldi (1951): *The principle of minimized iterations in the solution of the matrix eigenvalue problem*. Quart. Appl. Math., vol. 9, pp. 17-29. [IV.10]
- U. Ascher (1989): *On numerical differential algebraic problems with application to semiconductor device simulation*. SIAM J. Numer. Anal., vol. 26, pp. 517-538. [VII.4]
- U. Ascher & G. Bader (1986): *Stability of collocation at Gaussian points*. SIAM J. Numer. Anal., vol. 23, pp. 412-422. [IV.13]
- U.M. Ascher, H. Chin & S. Reich (1994): *Stabilization of DAEs and invariant manifolds*. Numer. Math., vol. 67, pp. 131-149. [VII.2]
- U. Ascher & L.R. Petzold (1991): *Projected implicit Runge-Kutta methods for differential-algebraic equations*. SIAM J. Numer. Anal., vol. 28, pp. 1097-1120. [VII.4]
- M. Athans & P.L. Falb (1966): *Optimal Control*. McGraw-Hill Book Company, New York, 879pp. [VII.1]
- W. Auzinger, R. Frank, & F. Macsek (1990): *Asymptotic error expansions for stiff equations: the implicit Euler scheme*. SIAM J. Numer. Anal., vol. 27, pp. 67-104. [VI.5]
- O. Axelsson (1969): *A class of A-stable methods*. BIT, vol. 9, pp. 185-199. [IV.3], [IV.5]
- O. Axelsson (1972): *A note on a class of strongly A-stable methods*. BIT, vol. 12, pp. 1-4. [IV.5]
- G. Bader & P. Deuflhard (1983): *A semi-implicit mid-point rule for stiff systems of ordinary differential equations*. Numer. Math., vol. 41, pp. 373-398. [IV.9], [IV.10], [VII.6]
- G. Bader, see also U. Ascher & G. Bader; E. Hairer, G. Bader & Ch. Lubich.
- C. Baiocchi & M. Crouzeix (1989): *On the equivalence of A-stability and G-stability*. Appl. Numer. Math., vol. 5, pp. 19-22. [V.6]
- M. Bakker (1971): *Analytical aspects of a minimax problem* (Dutch), Technical Note TN 62, Mathematical Centre, Amsterdam.
- L.A. Bales, O.A. Karakashian & S.M. Serbin (1988): *On the A_0 -acceptability of rational approximations to the exponential function with only real poles*. BIT, vol. 28, pp. 70-79. [IV.4]
- G.P. Barker, A. Berman & R.J. Plemmons (1978): *Positive diagonal solutions to the Lyapunov equations*. Linear and Multilinear Algebra, vol. 5, pp. 249-256. [IV.14]
- J. Baumgarte (1972): *Stabilization of constraints and integrals of motion in dynamical systems*. Comp. Meth. Appl. Mech. Eng., vol. 1, pp. 11-16. [VII.2]
- G. Benettin & A. Giorgilli (1994): *On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms*. J. Statist. Phys., vol. 74, pp. 1117-1143. [VII.8]
- H.J.C. Berendsen, see also J.-P. Ryckaert, G. Ciccotti & H.J.C. Berendsen.
- A. Berman, see G.P. Barker, A. Berman & R.J. Plemmons.
- S. Bernstein (1914): *Sur la définition et les propriétés des fonctions analytiques d'une variable réelle*. Math. Annalen, vol. 75, pp. 449-468. [IV.11]

- S. Bernstein (1928): *Sur les fonctions absolument monotones*. Acta Mathematica, vol. 51, pp. 1-66. [IV.11]
- M. Berzins & R.M. Furzeland (1985): *A user's manual for SPRINT – a versatile software package for solving systems of algebraic, ordinary and partial differential equations: part 1 – algebraic and ordinary differential equations*. Thornton Research Centre, Shell Research Ltd. TNER.85.058. [V.5], [VII.3]
- T.A. Bickart (1977): *An efficient solution process for implicit Runge-Kutta methods*. SIAM J. Numer. Anal., vol. 14, 1022-1027. [IV.8]
- T.A. Bickart & W.B. Rubin (1974): *Composite multistep methods and stiff stability*. In: Stiff Differential Systems, R.A. Willoughby (ed.), Plenum Press, New York. [V.3]
- T.A. Bickart, see also H.M. Sloate & T.A. Bickart.
- G. Birkhoff & R.S. Varga (1965): *Discretization errors for well-set Cauchy problems, I*. J. Math. Phys., vol. 44, pp. 1-23. [IV.5]
- Å. Björck (1983): *A block QR algorithm for partitioning stiff differential systems*. BIT, vol. 23, pp. 329-345. [IV.10]
- Å. Björck (1984): *Some methods for separating stiff components in initial value problems*. In: Numerical Analysis, Dundee 1983, D.F. Griffiths, ed., Lecture Notes in Math. 1066, Springer Verlag, pp. 30-43. [IV.10]
- C. Bolley & M. Crouzeix (1978): *Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques*. R.A.I.R.O. Analyse numérique, vol. 12, pp. 237-245. [IV.11]
- V.G. Boltyanskii, see L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze & E.F. Mishchenko.
- V. Brasey (1992): *A half-explicit Runge-Kutta method of order 5 for solving constrained mechanical systems*. Computing, vol. 48, pp. 191-201. [VII.6]
- V. Brasey (1994): *Half-explicit method for semi-explicit differential-algebraic equations of index 2*. Thèse N° 2664, Sect. Math., Univ. de Genève. [VII.7]
- V. Brasey & E. Hairer (1993): *Half-explicit Runge-Kutta methods for differential-algebraic systems of index 2*. SIAM J. Numer. Anal., vol. 30, pp. 538-552. [VII.6]
- K.E. Brenan (1983): *Stability and convergence of difference approximations for higher-index differential-algebraic systems with applications in trajectory control*. Doctoral thesis, Dep. Math., Univ. of California, Los Angeles. [VII.1]
- K.E. Brenan, S.L. Campbell & L.R. Petzold (1989): *Numerical solution of initial-value problems in differential-algebraic equations*. North Holland, New York, 210pp. [VII.1], [VII.3], [VII.7]
- K.E. Brenan & L.R. Engquist (1988): *Backward differentiation approximations of nonlinear differential/algebraic equations*, and Supplement. Math. Comp., vol. 51, pp. 659-676, pp. S7-S16. [VII.3]
- K.E. Brenan & L.R. Petzold (1989): *The numerical solution of higher index differential-/algebraic equations by implicit Runge-Kutta methods*. SIAM J. Numer. Anal., vol. 26, pp. 976-996. [VII.4]
- P.N. Brown, G.D. Byrne & A.C. Hindmarsh (1989): *VODE: a variable coefficient ODE solver*. SIAM J. Sci. Stat. Comput., vol. 10, pp. 1039-1051. [V.5]
- T.D. Bui, see P. Kaps, S.W.H. Poon & T.D. Bui.

- J.M. Burgers (1948): *A mathematical model illustrating the theory of turbulence*. Advances in appl. mech., vol. 1, pp. 171-199. [V.8], [VI.6]
- K. Burrage (1978): *High order algebraically stable Runge-Kutta methods*. BIT, vol. 18, pp. 373-383. [IV.5], [IV.13]
- K. Burrage (1978): *A special family of Runge-Kutta methods for solving stiff differential equations*. BIT, vol. 18, pp. 22-41. [IV.5], [IV.6], [IV.8]
- K. Burrage (1982): *Efficiently implementable algebraically stable Runge-Kutta methods*. SIAM J. Numer. Anal., vol. 19, pp. 245-258. [IV.13]
- K. Burrage (1987): *High order algebraically stable multistep Runge-Kutta methods*. SIAM J. Numer. Anal., vol. 24, pp. 106-115. [V.9]
- K. Burrage (1988): *Order properties of implicit multivalued methods for ordinary differential equations*. IMA J. Numer. Anal., vol. 8, pp. 43-69. [V.9]
- K. Burrage & J.C. Butcher (1979): *Stability criteria for implicit Runge-Kutta methods*. SIAM J. Numer. Anal., vol. 16, pp. 46-57. [IV.12]
- K. Burrage & J.C. Butcher (1980): *Non-linear stability of a general class of differential equation methods*. BIT, vol. 20, pp. 185-203. [IV.12], [V.9]
- K. Burrage, J.C. Butcher & F.H. Chipman (1980): *An implementation of singly-implicit Runge-Kutta methods*. BIT, vol. 20, pp. 326-340. [IV.8]
- K. Burrage & W.H. Hundsdorfer (1987): *The order of B-convergence of algebraically stable Runge-Kutta methods*. BIT, vol. 27, pp. 62-71. [IV.15]
- J.C. Butcher (1964): *Implicit Runge-Kutta processes*. Math. Comput., vol. 18, pp. 50-64. [IV.5]
- J.C. Butcher (1964): *Integration processes based on Radau quadrature formulas*. Math. Comput., vol. 18, pp. 233-244. [IV.5]
- J.C. Butcher (1975): *A stability property of implicit Runge-Kutta methods*. BIT, vol. 15, pp. 358-361. [IV.12]
- J.C. Butcher (1976): *On the implementation of implicit Runge-Kutta methods*. BIT, vol. 6, pp. 237-240. [IV.8]
- J.C. Butcher (1977): *On A-stable implicit Runge-Kutta methods*. BIT, vol. 17, pp. 375-378. [IV.5]
- J.C. Butcher (1979): *A transformed implicit Runge-Kutta method*. J. Assoc. Comput. Mach., vol. 26, pp. 731-738. [IV.8]
- J.C. Butcher (1981): *A generalization of singly-implicit methods*. BIT, vol. 21, pp. 175-189. [V.3]
- J.C. Butcher (1982): *A short proof concerning B-stability*. BIT, vol. 22, pp. 528-529. [IV.12]
- J.C. Butcher (1987): *Linear and non-linear stability for general linear methods*. BIT, vol. 27, pp. 182-189. [V.9]
- J.C. Butcher (1987): *The equivalence of algebraic stability and AN-stability*. BIT, vol. 27, pp. 510-533. [V.9]
- J.C. Butcher (1987): *The numerical analysis of ordinary differential equations. Runge-Kutta and general linear methods*. John Wiley & Sons, 512pp. [IV.12]
- J.C. Butcher (1990): *Order, stepsize and stiffness switching*. Computing, vol. 44, p. 209-220. [IV.2]

- J.C. Butcher, see also K. Burrage & J.C. Butcher; K. Burrage, J.C. Butcher & F.H. Chipman.
- G.D. Byrne & A.C. Hindmarsh (1975): *A polyalgorithm for the numerical solution of ordinary differential equations*. ACM Trans. Math. Software, vol. 1, pp. 71-96. [V.5]
- G.D. Byrne & A.C. Hindmarsh (1987): *Stiff ODE solvers: a review of current and coming attractions*. J. of Comput. Physics, vol. 70, pp. 1-62. [IV.10]
- G.D. Byrne, see also P.N. Brown, G.D. Byrne & A.C. Hindmarsh.
- D.A. Calahan (1968): *A stable, accurate method of numerical integration for nonlinear systems*. Proc. IEEE, vol. 56, p. 744. [IV.7]
- A. Callender, D.R. Hartree & A. Porter (1936): *Time-lag in a control system*. Phil. Trans. of the Royal Society (London), Series A, vol. 235, pp. 415-444. [IV.2]
- M.P. Calvo, see also J.M. Sanz-Serna & M.P. Calvo.
- S.L. Campbell (1982): *Singular Systems of Differential Equations II*. Pitman, London. [VII.1]
- S.L. Campbell (1989): *A computational method for general higher index singular systems of differential equations*. IMACS Transactions Scientific Computing, vol. 1.2, pp. 555-560. [VII.2]
- S.L. Campbell (1993): *Least squares completions for nonlinear differential algebraic equations*. Numer. Math., vol. 65, pp. 77-94. [VII.2]
- S.L. Campbell (1995): *High index differential algebraic equations*. J. Mech. Struct. & Machines, vol. 23, pp. 199-222. [VII.1]
- S.L. Campbell & C.W. Gear (1995): *The index of general nonlinear DAEs*. Numer. Math., vol. 72, pp. 173-196. [VII.1]
- S.L. Campbell & E. Moore (1995): *Constraint preserving integrators for general nonlinear higher index DAEs*. Numer. Math., vol. 69, pp. 383-399. [VII.2]
- S.L. Campbell, see also K.E. Brenan, S.L. Campbell & L.R. Petzold.
- J. Carr, D.B. Duncan & C.H. Walshaw (1995): *Numerical approximation of a metastable system*. IMA J. Numer. Anal., vol. 15, pp. 505-521. [IV.10]
- J.R. Cash (1976): *Semi-implicit Runge-Kutta procedures with error estimates for the numerical integration of stiff systems of ordinary differential equations*. JACM, vol. 23, pp. 455-460. [IV.7]
- J.R. Cash (1979): *Diagonally implicit Runge-Kutta formulae with error estimates*. J. Inst. Math. Applics, vol. 24, pp. 293-301. [IV.6]
- J.R. Cash (1979): *Stable recursions, with applications to the numerical solution of stiff systems*. Academic Press, 223 pp. [V.2]
- J.R. Cash (1980): *On the integration of stiff systems of O.D.E.s using extended backward differentiation formulae*. Numer. Math., vol. 34, pp. 235-246. [V.3]
- J.R. Cash (1981): *Second derivative extended backward differentiation formulas for the numerical integration of stiff systems*. SIAM J. Numer. Anal. vol. 18, pp. 21-36. [V.3]
- J.R. Cash (1983): *The integration of stiff initial value problems in ODEs using modified extended backward differentiation formulas*. Comp. & Maths. with Appls., vol. 9, No. 5, pp. 645-657. [V.3], [V.5]
- J.R. Cash & S. Considine (1992): *An MEBDF code for stiff initial value problems*. ACM Trans. Math. Software, vol. 18, No. 2, pp. 142-158. [V.5]

- P.E. Chase (1962): *Stability properties of Predictor-Corrector methods for ordinary differential equations*, J. Assoc. Comput. Mach., vol. 9, pp.457-468. [V.1]
- P.L. Chebyshev (Tchébychef) (1854): *Théorie des mécanismes connus sous le nom de parallélogrammes*. Mém. de l'Acad. Imp. St.-Petersbourg, tome VII (1854), pp.539-568; Oeuvres Tome I, pp.111-143. [IV.2]
- H. Chin, see also U.M. Ascher, H. Chin & S. Reich.
- F.H. Chipman (1971): *A-stable Runge-Kutta processes*. BIT, vol. 11, pp. 384-388. [IV.5]
- F.H. Chipman (1976): *A note on implicit A-stable RK methods with parameters*. BIT, vol. 16, pp. 223-227. [IV.5]
- F.H. Chipman, see also K. Burrage, J.C. Butcher & F.H. Chipman.
- G. Ciccotti, see also J.-P. Ryckaert, G. Ciccotti & H.J.C. Berendsen.
- K. Clark (1988): *A structural form for higher index semistate equations I: Theory and applications to circuit and control theory*. Linear Alg. Appl., vol. 98, pp. 169-197. [VII.1]
- L. Collatz (1950): *Numerische Behandlung von Differentialgleichungen*. Grundlehren, Springer Verlag, Band LX (later editions and translations). [IV.10], [IV.15]
- P. Collet, J.-P. Eckmann, H. Epstein & J. Stubbe (1993): *Analyticity for the Kuramoto-Sivashinsky equation*. Physica D, vol. 67, pp. 321-326. [IV.10]
- S. Considine, see also J.R. Cash & S. Considine.
- G.J. Cooper (1985): *Reducible Runge-Kutta methods*. BIT, vol. 25, pp. 675-680. [IV.12]
- G.J. Cooper (1986): *On the existence of solutions for algebraically stable Runge-Kutta methods*. IMA J. Numer. Anal., vol. 6, pp. 325-330. [IV.14]
- G.J. Cooper & A. Sayfy (1979): *Semiexplicit A-stable Runge-Kutta methods*. Math. of Comp., vol. 33, pp. 541-556. [IV.6]
- G.J. Cooper & A. Sayfy (1983): *Additive Runge-Kutta methods for stiff ordinary differential equations*. Math. of Comp., vol. 40, pp. 207-218. [IV.7]
- R. Courant, K. Friedrichs & H. Lewy (1928): *Ueber die partiellen Differenzengleichungen der mathematischen Physik*. Math. Ann., vol. 100, pp. 32-74. [IV.2]
- R. Courant, see A. Hurwitz & R. Courant.
- G. Cramer (1750): *Introduction à l'analyse des lignes courbes algébriques*. Genève, 1750. [IV.3]
- R.L. Crane & R.W. Klopfenstein (1965): *A predictor-corrector algorithm with an increased range of absolute stability*. J. ACM, vol. 12, pp.227-241. . [V.1]
- M. Crouzeix (1975): *Sur l'approximation des équations différentielles opérationnelles linéaires par de méthodes de Runge-Kutta*. Thèse, Univ. Paris VI. [IV.6]
- M. Crouzeix (1979): *Sur la B-stabilité des méthodes de Runge-Kutta*. Numer. Math., vol. 32, pp. 75-82. [IV.12]
- M. Crouzeix, W.H. Hundsdorfer & M.N. Spijker (1983): *On the existence of solutions to the algebraic equations in implicit Runge-Kutta methods*. BIT, vol. 23, pp. 84-91. [IV.14]
- M. Crouzeix & P.A. Raviart (1976): *Approximation des équations d'évolution linéaires par des méthodes à pas multiples*. C. R. Acad. Sc. Paris, Ser. A 283, pp. 367-370. [V.7]
- M. Crouzeix & P.A. Raviart (1980): *Approximation des problèmes d'évolution*. Unpublished Lecture Notes, Université de Rennes. [IV.6], [IV.14], [V.7]

- M. Crouzeix & F. Ruamps (1977): *On rational approximations to the exponential*. R.A.I.R.O. Analyse Numérique, vol. 11, pp. 241-243. [IV.4]
- M. Crouzeix, see also C. Baiocchi & M. Crouzeix; C. Bolley & M. Crouzeix.
- C.W. Cryer (1973): *A new class of highly stable methods*. A_0 -stable methods. BIT, vol. 13, pp. 153-159. [V.2]
- A.R. Curtis (1983): *Jacobian matrix properties and their impact on choice of software for stiff ODE systems*. IMA J. Numer. Anal., vol. 3, pp. 397-415. [IV.10]
- C.F. Curtiss & J.O. Hirschfelder (1952): *Integration of stiff equations*. Proc. Nat. Acad. Sci., vol. 38, pp. 235-243. [IV.1]
- G. Dahlquist (1951): *Fehlerabschätzungen bei Differenzenmethoden zur numerischen Integration gewöhnlicher Differentialgleichungen*. ZAMM, vol. 31, pp. 239-240. [V.1]
- G. Dahlquist (1956): *Convergence and stability in the numerical integration of ordinary differential equations*. Math. Scand., vol. 4, pp. 33-53. [V.7]
- G. Dahlquist (1963): *A special stability problem for linear multistep methods*. BIT, vol. 3, pp. 27-43. [IV.3], [IV.9], [IV.12], [V.1], [V.6]
- G. Dahlquist (1975): *Error analysis for a class of methods for stiff nonlinear initial value problems*. Numerical Analysis, Dundee 1975, Lecture Notes in Math., No. 506, pp. 60-74. [IV.12], [V.6]
- G. Dahlquist (1978): *G-stability is equivalent to A-stability*. BIT, vol. 18, pp. 384-401. [IV.13], [V.6]
- G. Dahlquist (1978): *Positive functions and some applications to stability questions for numerical methods*. In: Recent Advances in Numerical Analysis, C. de Boor & G.H. Golub (eds.), Academic Press, New York, pp. 1-19. [IV.5]
- G. Dahlquist (1983): *On one-leg multistep methods*. SIAM J. Numer. Anal., vol. 20, pp. 1130-1138. [V.6], [V.7], [V.9]
- G. Dahlquist & R. Jeltsch (1979): *Generalized disks of contractivity for explicit and implicit Runge-Kutta methods*. TRITA-NA Report 7906. [IV.12], [IV.13]
- G. Dahlquist & R. Jeltsch (1987): *Reducibility and contractivity of Runge-Kutta methods revisited*. Report Nr. 46, Inst. f. Geometrie u. Prakt. Math., RWTH Aachen. [IV.12]
- G. Dahlquist, H. Mingyou & R. LeVeque (1983): *On the uniform power-boundedness of a family of matrices and the applications to one-leg and linear multistep methods*. Numer. Math., vol. 42, pp. 1-13. [V.7]
- G. Dahlquist & G. Söderlind (1982): *Some problems related to stiff nonlinear differential systems*. In: Computing Methods in Applied Sciences and Engineering, V.R. Glowinski & J.L. Lions (eds.), North-Holland, INRIA [V.7]
- G. Dahlquist, see also G. Söderlind & G. Dahlquist.
- J.W. Daniel & R.E. Moore (1970): *Computation and theory in ordinary differential equations*, W.H. Freeman and Company, 172 pp. [V.4]
- P.J. Davis (1963): *Interpolation and approximation*. Blaisdell 1963; Dover 1975. [V.3]
- K. Dekker (1981): *Stability of linear multistep methods on the imaginary axis*. BIT, vol. 21, pp. 66-79. [V.4]
- K. Dekker (1982): *On the iteration error in algebraically stable Runge-Kutta methods*. Report NW 138/82, Math. Centrum, Amsterdam. [IV.14]

- K. Dekker (1984): *Error bounds for the solution to the algebraic equations in Runge-Kutta methods*. BIT, vol. 24, pp. 347-356. [IV.14]
- K. Dekker & E. Hairer (1985): *A necessary condition for BSI-stability*. BIT, vol. 25, pp. 285-288. [IV.14]
- K. Dekker, J.F.B.M. Kraaijevanger & J. Schneid (1990): *On the relation between algebraic stability and B-convergence for Runge-Kutta methods*. Numer. Math., vol. 57, pp. 249-262. [IV.15]
- K. Dekker & J.G. Verwer (1984): *Stability of Runge-Kutta methods for stiff nonlinear differential equations*. North-Holland, Amsterdam-New-York-Oxford. [IV.12], [IV.14], [IV.15]
- K. Dekker, see also M.Z. Liu, K. Dekker & M.N. Spijker.
- P. Deuflhard (1983): *Order and stepsize control in extrapolation methods*. Numer. Math., vol. 41, pp. 399-422. [IV.9]
- P. Deuflhard (1985): *Recent progress in extrapolation methods for ordinary differential equations*. SIAM Review, vol. 27, pp. 505-535. [IV.9]
- P. Deuflhard, E. Hairer & J. Zugck (1987): *One-step and extrapolation methods for differential-algebraic systems*. Numer. Math., vol. 51, pp. 501-516. [VI.5]
- P. Deuflhard & U. Nowak (1987): *Extrapolation integrators for quasilinear implicit ODEs*. In P. Deuflhard & B. Engquist (eds.), *Large-Scale Scientific Computing*. Birkhäuser, Boston. [VI.5], [VI.6]
- P. Deuflhard, see also G. Bader & P. Deuflhard.
- G.A. Di Marzo (1992): *RODAS5(4), méthodes de Rosenbrock d'ordre 5(4) adaptées aux problèmes différentiels-algébriques*. Mémoire de diplôme en Mathématiques, Université de Genève 1992. [IV.10], [VI.4]
- J.R. Dormand & P.J. Prince (1980): *A family of embedded Runge-Kutta formulae*. J. Comp. Appl. Math., vol. 6, pp. 19-26. [IV.2]
- A.A. Dorodnicyn (1947): *Asymptotic solution of the van der Pol equation*. Prikl. Mat. i Meh., vol. 11, pp. 313-328; Translations AMS, Ser. 1, vol. 4, pp. 1-23. [VI.1]
- B.L. Ehle (1968): *High order A-stable methods for the numerical solution of systems of DEs*. BIT, vol. 8, pp. 276-278. [IV.3], [IV.4], [IV.5]
- B.L. Ehle (1969): *On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems*. Research Report CSRR 2010, Dept. AACS, Univ. of Waterloo, Ontario, Canada. [IV.3], [IV.5]
- B.L. Ehle (1973): *A-stable methods and Padé approximations to the exponential*. SIAM J. Math. Anal., vol. 4, pp. 671-680. [IV.4], [IV.5]
- B.L. Ehle & Z. Picel (1975): *Two-parameter, arbitrary order, exponential approximations for stiff equations*. Math. Comput., vol. 29, pp. 501-511. [IV.5]
- E. Eich (1993): *Convergence results for a coordinate projection method applied to mechanical systems with algebraic constraints*. SIAM J. Numer. Anal., vol. 30, pp. 1467-1482. [VII.2]
- R. England (1982): *Some hybrid implicit stiffly stable methods for ordinary differential equations*. In: Numerical Analysis, Proc. Mexico, (ed. J.P. Hennart), Lecture Notes in Math., No. 909, Springer Verlag, pp. 147-158. [V.3]
- L.R. Engquist, see K.E. Brenan & L.R. Engquist.

- W.H. Enright (1974): *Optimal second derivative methods for stiff systems*. In: Stiff Differential Systems, ed. by R.A. Willoughby, Plenum Press, New York. [V.3]
- W.H. Enright (1974): *Second derivative multistep methods for stiff ordinary differential equations*, SIAM J. Numer. Anal., vol. 11, pp. 321-331. [V.3]
- W.H. Enright (1978): *Improving the efficiency of matrix operations in the numerical solution of stiff ordinary differential equations*. ACM Trans. on Math. Software, vol. 4, pp. 127-136. [IV.8]
- W.H. Enright & T.E. Hull (1976): *Comparing numerical methods for the solution of stiff systems of ODEs arising in chemistry*. In: Numerical methods for differential systems, recent developments in algorithms, software and applications, L. Lapidus & W.E. Schiesser, Eds., Academic Press, New York, 1976, pp. 45-66. [IV.10]
- W.H. Enright, T.E. Hull & B. Lindberg (1975): *Comparing numerical methods for stiff systems of ODEs*. BIT, vol. 15, pp. 10-48. [IV.10]
- W.H. Enright & M.S. Kamel (1979): *Automatic partitioning of stiff systems and exploiting the resulting structure*. ACM TOMS, vol. 5, pp. 374-385. [IV.10]
- M.A. Epton, see R.F. Sincovec, A.M. Erisman, E.L. Yip & M.A. Epton.
- A.M. Erisman, see R.F. Sincovec, A.M. Erisman, E.L. Yip & M.A. Epton.
- L. Euler (1737): *De fractionibus continuis dissertatio*. Comm. acad. sc. Petrop., vol. 9, pp. 98-137; Opera Omnia vol. XIV, pp. 187-215 (vide §7). [IV.13]
- L. Euler (1752): *Elementa doctrinae solidorum*. Nov. comm. acad. sci. Petropolitanae vol. 4, p. 109-140; Opera Omnia vol. XXVI, pp. 71-93. [IV.4]
- P.L. Falb, see M. Athans & P.L. Falb.
- L. Fejér (1933): *Mechanische Quadraturen mit positiven Coteschen Zahlen*. Math. Zeitschrift, vol. 37, pp. 287-309. [IV.13]
- A. Feng, C.D. Holland & S.E. Gallun (1984): *Development and comparison of a generalized semi-implicit Runge-Kutta method with Gear's method for systems of coupled differential and algebraic equations*. Comp. & Chem. Eng., vol. 8, pp. 51-59. [VI.4]
- J. Field & R.M. Noyes (1974): *Oscillations in chemical systems. IV: Limit cycle behavior in a model of a real chemical reaction*. J. Chem. Phys., vol. 60, pp. 1877-1884. [IV.10]
- R. Frank, J. Schneid & C.W. Ueberhuber (1981): *The concept of B-convergence*. SIAM J. Numer. Anal., vol. 18, pp. 753-780. [IV.15]
- R. Frank, J. Schneid & C.W. Ueberhuber (1985): *Stability properties of implicit Runge-Kutta methods*. SIAM J. Numer. Anal., vol. 22, pp. 497-514. [IV.14], [IV.15]
- R. Frank, J. Schneid & C.W. Ueberhuber (1985): *Order results for implicit Runge-Kutta methods applied to stiff systems*. SIAM J. Numer. Anal., vol. 22, pp. 515-534. [IV.14], [IV.15]
- R. Frank, see also W. Auzinger, R. Frank, & F. Macsek.
- J.N. Franklin (1959): *Numerical stability in digital and analogue computation for diffusion problems*. J. Math. Phys., vol 37, pp. 305-315. [IV.2]
- A. Friedli (1978): *Verallgemeinerte Runge-Kutta Verfahren zur Lösung steifer Differentialgleichungssysteme*. Oberwolfach Conference 1976, Lecture Notes in Math. 631, pp. 35-50. [IV.11]
- K. Friedrichs, see R. Courant, K. Friedrichs & H. Lewy.

- C. Führer (1988): *Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen: Theorie, numerische Ansätze und Anwendungen*. Doctoral thesis, Technische Universität München [VII.2].
- C. Führer & B.J. Leimkuhler (1991): *Numerical solution of differential-algebraic equations for constrained mechanical motion*. Numer. Math., vol. 59, pp. 55-69. [VII.2]
- C. Führer, see also C. Arévalo, C. Führer & G. Söderlind.
- H. Fujita & T. Kato (1964): *On the Navier-Stokes initial value problem. I*. Arch. Rat. Mech. Anal., vol. 16, pp. 269-315. [V.8]
- R.M. Furzeland, see M. Berzins & R.M. Furzeland.
- B.G. Galerkin (1915): *Series expansions for some cases of equilibria of plates and beams* (Russian). Vestnik Ingenerov Petrograd, H.10. [IV.10]
- S.E. Gallun, see A. Feng, C.D. Holland & S.E. Gallun.
- R.V. Gamkrelidze, see L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze & E.F. Mishchenko.
- F.R. Gantmacher (1954): *Teorya Matrits*. Two volumes, Gosudarstv. Izdat. Techn.-Teor. Lit., Moscva 1953; translations: Chelsea NY 1959, Interscience NY and London 1959, D. Verl. d. Wiss. Berlin 1958/59, Dunod Paris 1966. [VII.1]
- C.W. Gear (1971): *Numerical initial value problems in ordinary differential equations*, Prentice Hall, 253 pp. [V.2], [V.5]
- C.W. Gear (1971): *Simultaneous numerical solution of differential-algebraic equations*. IEEE Trans. Circuit Theory, vol. CT-18, pp. 89-95. [VI.2]
- C.W. Gear (1982): *Automatic detection and treatment of oscillatory and/or stiff ordinary differential equations*. In: Numerical integration of differential equations, Lecture Notes in Math., vol. 968, pp. 190-206. [IV.1]
- C.W. Gear (1988): *Differential-algebraic equation index transformations*. SIAM J. Sci. Stat. Comput., vol. 9, pp. 39-47. [VII.4]
- C.W. Gear (1990): *Differential-algebraic equations, indices, and integral algebraic equations*. SIAM J. Numer. Anal., vol. 27. [VII.1]
- C.W. Gear, G.K. Gupta & B. Leimkuhler (1985): *Automatic integration of Euler-Lagrange equations with constraints*. J. Comp. Appl. Math., vol. 12 & 13, pp. 77-90. [VII.1], [VII.3], [VII.7]
- C.W. Gear, H.H. Hsu & L. Petzold (1981): *Differential-algebraic equations revisited*. Proc. Numerical Methods for Solving Stiff Initial Value Problems, Oberwolfach, BRD. [VII.3]
- C.W. Gear & L.R. Petzold (1983): *Differential/algebraic systems and matrix pencils*. In: Matrix Pencils, B. Kagstrom & A. Ruhe (eds.), Lecture Notes in Math. 973, Springer Verlag, pp. 75-89. [VII.1]
- C.W. Gear & L.R. Petzold (1984): *ODE methods for the solution of differential/algebraic systems*. SIAM J. Numer. Anal., vol. 21, pp. 716-728. [VII.1], [VII.3]
- C.W. Gear & Y. Saad (1983): *Iterative solution of linear equations in ODE codes*. SIAM J. Sci. Stat. Comput., vol. 4, pp. 583-601. [IV.10]
- C.W. Gear, see also S.L. Campbell & C.W. Gear.
- E. Gekeler (1979): *Uniform stability of linear multistep methods in Galerkin procedures for parabolic problems*. J. Math. Sciences, vol. 2, pp. 651-667. [V.7]

- E. Gekeler (1984): *Discretization Methods for Stable Initial Value Problems*. Lecture Notes in Math., No. 1044, Springer Verlag. [V.7]
- Y. Genin (1974): *An algebraic approach to A-stable linear multistep-multiderivative integration formulas*. BIT, vol. 14, pp. 382-406. [V.4]
- D.R.A. Giles (1978): *A comparison of three problem-oriented simulation programs for dynamic mechanical systems*. Thesis, Univ. Waterloo, Ontario. [VII.7]
- A. Giorgilli, see also G. Benettin & A. Giorgilli.
- G.H. Golub & C.F. Van Loan (1989): *Matrix Computations*. Second edition, John Hopkins Univ. Press, Baltimore and London. [VII.1]
- B.A. Gottwald (1977): *MISS — Ein einfaches Simulations-System für biologische und chemische Prozesse*, EDV in Medizin und Biologie, vol. 3, pp. 85-90. [IV.10]
- A.R. Gourlay (1970): *A note on trapezoidal methods for the solution of initial value problems*. Math. of Comp., vol. 24, pp. 629-633. [IV.3]
- J.A. van de Griend & J.F.B.M. Kraaijevanger (1986): *Absolute monotonicity of rational functions occurring in the numerical study of initial value problems*. Numer. Math., vol. 49, pp. 413-424. [IV.11]
- E. Griepentrog & R. März (1986): *Differential-algebraic equations and their numerical treatment*. Teubner Texte zur Math., Band 88. [VI.1], [VII.1], [VII.3]
- R.D. Grigorieff (1977): *Numerik gewöhnlicher Differentialgleichungen, Bd. 2, Mehrschrittverfahren*. Teubner Studienbücher, 411 Seiten "mit 49 Figuren, 32 Tabellen und zahlreichen Beispielen". [V.1]
- R.D. Grigorieff & J. Schroll (1978): *Über $A(\alpha)$ -stabile Verfahren hoher Konsistenzordnung*. Computing, vol. 20, pp. 343-350. [V.2]
- A. Guillou & B. Lago (1961): *Domaine de stabilité associé aux formules d'intégration numérique d'équations différentielles, à pas séparés et à pas liés. Recherche de formules à grand rayon de stabilité*. 1er Congr. Assoc. Fran. Calcul, AFCAL, Grenoble, Sept. 1960, pp. 43-56. [IV.2]
- A. Guillou & J.L. Soulé (1969): *La résolution numérique des problèmes différentiels aux conditions initiales par des méthodes de collocation*. R.I.R.O., vol. R-3, pp. 17-44. [V.3]
- G.K. Gupta, see C.W. Gear, G.K. Gupta & B. Leimkuhler.
- K. Gustafsson (1991): *Control theoretic techniques for stepsize selection in explicit Runge-Kutta methods*. ACM Trans. Math. Soft., vol. 17, pp. 533-554. [IV.2]
- K. Gustafsson (1994): *Control-theoretic techniques for stepsize selection in implicit Runge-Kutta methods*. ACM Trans. Math. Soft., vol. 20, pp. 496-517. [IV.8]
- K. Gustafsson, M. Lundh & G. Söderlind (1988): *A PI stepsize control for the numerical solution of ordinary differential equations*. BIT, vol. 28, pp. 270-287. [IV.2]
- E. Hairer (1980): *Highest possible order of algebraically stable diagonally implicit Runge-Kutta methods*. BIT, vol. 20, pp. 254-256. [IV.13]
- E. Hairer (1982): *Constructive characterization of A-stable approximations to $\exp z$ and its connection with algebraically stable Runge-Kutta methods*. Numer. Math., vol. 39, pp. 247-258. [IV.5]
- E. Hairer (1986): *A- and B-stability for Runge-Kutta methods - characterizations and equivalence*. Numer. Math., vol. 48, pp. 383-389. [IV.13]

- E. Hairer (1994): *Backward analysis of numerical integrators and symplectic methods*. Annals of Numer. Math., vol. 1, pp. 107-132. [VII.8]
- E. Hairer, G. Bader & Ch. Lubich (1982): *On the stability of semi-implicit methods for ordinary differential equations*. BIT, vol. 22, pp. 211-232. [IV.9], [IV.11]
- E. Hairer & Ch. Lubich (1988): *Extrapolation at stiff differential equations*. Numer. Math., vol. 52, pp. 377-400. [VI.5]
- E. Hairer & Ch. Lubich (1988b): *On extrapolation methods for stiff and differential-algebraic equations*. Teubner Texte zur Mathematik, Band 104, Teubner, Leipzig, pp. 64-73. [VI.5]
- E. Hairer & Ch. Lubich (1996): *The life-span of backward error analysis for numerical integrators*. Numer. Math. [VII.8]
- E. Hairer, Ch. Lubich & M. Roche (1988): *Error of Runge-Kutta methods for stiff problems studied via differential algebraic equations*. BIT, vol. 28, pp. 678-700. [VI.3]
- E. Hairer, Ch. Lubich & M. Roche (1989): *Error of Rosenbrock methods for stiff problems studied via differential algebraic equations*. BIT, vol. 29, pp. 77-90. [VI.3]
- E. Hairer, Ch. Lubich & M. Roche (1989): *The numerical solution of differential-algebraic systems by Runge-Kutta methods* (abbreviated as HLR89). Lecture Notes in Math. 1409, Springer Verlag. [VI.1], [VII.1], [VII.3], [VII.4], [VII.5], [VII.7]
- E. Hairer & A. Ostermann (1990): *Dense output for extrapolation methods*. Numer. Math., vol. 58, pp. 419-439. [VI.5]
- E. Hairer & H. Türke (1984): *The equivalence of B-stability and A-stability*. BIT, vol. 24, pp. 520-528. [IV.5], [IV.13]
- E. Hairer & G. Wanner (1981): *Algebraically stable and implementable Runge-Kutta methods of high order*. SIAM J. Numer. Anal., vol. 18, pp. 1098-1108. [IV.5], [IV.13]
- E. Hairer & G. Wanner (1982): *Characterization of non-linearly stable implicit Runge-Kutta methods*. In: Numerical integration of differential equations, Lecture Notes in Math., vol. 968, pp. 207-219. [IV.5], [IV.13]
- E. Hairer & G. Wanner (1995): *Analysis by its history*. Undergraduate Texts in Mathematics, Springer-Verlag New York. [IV.4],
- E. Hairer & G. Wanner (1996): *On a generalization of a theorem of von Neumann*. To appear in ZAMM. [IV.12]
- E. Hairer & M. Zennaro (1996): *On error growth functions of Runge-Kutta methods*. To appear in Appl. Numer. Math. [IV.11], [IV.12]
- E. Hairer, see also V. Brasey & E. Hairer; K. Dekker & E. Hairer; P. Deuffhard, E. Hairer & J. Zugck; G. Wanner, E. Hairer & S.P. Nørsett.
- G. Hall (1985): *Equilibrium states of Runge-Kutta schemes*. ACM Trans. Math. Software, vol. 11, pp. 289-301. [IV.1], [IV.2]
- G. Hall (1986): *Equilibrium states of Runge-Kutta schemes, part II*. ACM Trans. Math. Software, vol. 12, pp. 183-192. [IV.2]
- G. Hall & D.J. Higham (1988): *Analysis of stepsize selection schemes for Runge-Kutta codes*. IMA J. Numer. Anal., vol. 8, pp. 305-310. [IV.2]
- G. Hall, see also D.J. Higham & G. Hall.
- R.W.Hamming (1959): *Stable predictor-corrector methods for ordinary differential equations*. J. ACM, vol. 6, pp. 37-47. [V.1]

- R.W. HansonSmith, see D.S. Watkins & R.W. HansonSmith.
- D.R. Hartree, see A. Callender, D.R. Hartree & A. Porter.
- E.J. Haug (1989): *Computer-aided Kinematics and Dynamics of Mechanical Systems*. Allyn & Bacon, Boston. [VII.7]
- E.J. Haug, see also R.A. Wehage & E.J. Haug.
- F. Hausdorff (1921): *Summationsmethoden und Momentfolgen*. Math. Zeitschrift, vol. 9, pp. 74-109 and pp. 280-299. [IV.11]
- P. Henrici (1962): *Discrete Variable Methods in Ordinary Differential Equations*. Wiley, New York. [V.7]
- D. Henry (1981): *Geometric Theory of Semilinear Parabolic Equations*. Springer Lecture Notes in Mathematics 840. [V.8]
- Ch. Hermite (1873): *Sur la fonction exponentielle*. Comptes rendus de l'Acad. Sciences, vol. 77, pp. 18-24, 74-79, 226-233, 285-293. Œuvres, tome III, pp. 150-181. [IV.3]
- K.L. Hiebert, see L.F. Shampine & K.L. Hiebert.
- D.J. Higham & G. Hall (1990): *Embedded Runge-Kutta formulae with stable equilibrium states*. J. of Comp. and Appl. Math., vol. 29, pp. 25-33. [IV.2]
- D.J. Higham (1989): *Analysis of the Enright-Kamel partitioning method for stiff ordinary differential equations*. IMA J. Numer. Anal., vol. 9, pp. 1-14. [IV.10]
- D.J. Higham, see also G. Hall & D.J. Higham.
- A.C. Hindmarsh (1980): *LSODE and LSODI, two new initial value ordinary differential equation solvers*. ACM-SIGNUM Newsletter 15, pp. 10-11. [IV.10], [V.5], [VII.3]
- A.C. Hindmarsh (1983): *ODEPACK, a systematized collection of ode solvers*. In Scientific Computing, R.S. Stepleman et al. (eds.), North-Holland, Amsterdam, pp. 55-64. [V.5]
- A.C. Hindmarsh, see also P.N. Brown, G.D. Byrne & A.C. Hindmarsh; G.D. Byrne & A.C. Hindmarsh.
- J.O. Hirschfelder, see C.F. Curtiss & J.O. Hirschfelder.
- E. Hofer (1976): *A partially implicit method for large stiff systems of ODE's with only few equations introducing small time-constants*. SIAM J. Numer. Anal., vol. 13, pp. 645-663. [IV.10]
- C.D. Holland, see A. Feng, C.D. Holland & S.E. Gallun.
- E. Hopf (1950): *The partial differential equation $u_t + uu_x = \mu u_{xx}$* . Comm. on Pure and Appl. Math., vol. 3, pp. 201-230. [VI.5], [VI.6]
- P.J. van der Houwen (1968): *Finite difference methods for solving partial differential equations*. MC Tract 20, Math. Centrum, Amsterdam. [IV.2]
- P.J. van der Houwen (1973): *One-step methods with adaptive stability functions for the integration of differential equations*. Lecture Notes in Mathematics No. 333, Springer-Verlag, Berlin, pp. 164-174. [IV.7]
- P.J. van der Houwen (1977): *Construction of integration formulas for initial value problems*. North Holland series in Applied Math. and Mech., 269 pp. [IV.2], [IV.11]
- P.J. van der Houwen & B.P. Sommeijer (1980): *On the internal stability of explicit, m -stage Runge-Kutta methods for large m -values*. Z. Angew. Math. Mech., vol. 60, pp. 479-485. [IV.2]

- H.H. Hsu, see C.W. Gear, H.H. Hsu & L. Petzold.
- T.E. Hull, see W.H. Enright & T.E. Hull; W.H. Enright, T.E. Hull & B. Lindberg.
- W.H. Hundsdorfer (1985): *The numerical solution of nonlinear stiff initial value problems: an analysis of one step methods*. CWI Tract, Nr. 12, Mathematisch Centrum, Amsterdam. [IV.11], [IV.12], [IV.14]
- W.H. Hundsdorfer (1986): *Stability and B-convergence of linearly implicit Runge-Kutta methods*. Numer. Math., vol. 50, pp. 83-95. [IV.15]
- W.H. Hundsdorfer & M.N. Spijker (1981): *A note on B-stability of Runge-Kutta methods*. Numer. Math., vol. 36, pp. 319-331. [IV.12]
- W.H. Hundsdorfer & M.N. Spijker (1987): *On the algebraic equations in implicit Runge-Kutta methods*. SIAM J. Numer. Anal., vol. 24, pp. 583-594. [IV.14]
- W.H. Hundsdorfer & B.I. Steininger (1991): *Convergence of linear multistep and one-leg methods for stiff nonlinear initial value problems*. BIT vol. 31, p.124-143. [V.6], [V.7]
- W.H. Hundsdorfer, see also K. Burrage & W.H. Hundsdorfer; M. Crouzeix, W.H. Hundsdorfer & M.N. Spijker; J.G. Verwer, W.H. Hundsdorfer & B.P. Sommeijer.
- A. Hurwitz & R. Courant (1925): *Funktionentheorie*. 2. Aufl., Verlag von Julius Springer, Berlin. [V.4]
- A.F. Huxley, see A.L. Hodgkin & A.F. Huxley.
- A. Iserles (1981): *Generalized order star theory, in : Padé approximations and its applications*. Amsterdam 1980, ed. M.G. de Bruin & H. van Rossum, Lecture Notes in Math. #888. [IV.4]
- A. Iserles & S.P. Nørsett (1984): *A proof of the first Dahlquist barrier by order stars*. BIT, vol. 24, pp. 529-537. [V.4]
- A. Iserles & G. Strang (1983): *The optimal accuracy of difference schemes*. Trans. Am. Math. Soc., vol. 277, pp. 779-803. [IV.4]
- A. Iserles & R.A. Williamson (1983): *Stability and accuracy of semi-discretized finite difference methods*. IMA J. Numer. Anal., vol. 4, pp. 289-307. [IV.4]
- C.G.J. Jacobi (1826): *Ueber Gauss' neue Methode die Werthe der Integrale näherungsweise zu finden*. Journ. f. reine u. angew. Math., vol. I, pp. 301-308; Werke Vol. VI (1981), pp. 1-11. [IV.5]
- L. Jay (1993): *Convergence of a class of Runge-Kutta methods for differential-algebraic systems of index 2*, BIT, vol. 33, pp. 137-150. [VII.4]
- L. Jay (1994): *Runge-Kutta type methods for index three differential-algebraic equations with applications to Hamiltonian systems*. Thesis No. 2658, Univ. Genève. [VII.8]
- L. Jay (1995): *Structure-preserving integrators*. Submitted for publication. [VII.8]
- L. Jay (1996): *Symplectic partitioned Runge-Kutta methods for constrained Hamiltonian systems*. SIAM J. Numer. Anal., vol. 33, pp. 368-387. [VII.8]
- R. Jeltsch (1976): *Stiff stability and its relation to A_0 - and $A(0)$ -stability*, SIAM J. Numer. Anal., vol. 13, pp. 8-17. [V.2]
- R. Jeltsch (1976): *Note on A-stability of multistep multiderivative methods*. BIT, vol. 16, pp. 74-78. [V.4]
- R. Jeltsch (1978): *Stability on the imaginary axis and A-stability of linear multistep methods*. BIT, vol. 18, pp. 170-174. [V.4]

- R. Jeltsch (1988): *Order barriers for difference schemes for linear and nonlinear hyperbolic problems*. In: Numerical Analysis 1987, D.F. Griffiths & G.A. Watson (eds.), Pitman Research Notes in Math., No. 170, pp. 157-175. [IV.4]
- R. Jeltsch & O. Nevanlinna (1978): *Largest disk of stability of explicit Runge-Kutta methods*. BIT, vol. 18, pp. 500-502. [IV.4]
- R. Jeltsch & O. Nevanlinna (1981): *Stability of explicit time discretizations for solving initial value problems*. Numer. Math., vol. 37, pp. 61-91; Corrigendum: Numer. Math., vol. 39, p.155. [IV.4]
- R. Jeltsch & O. Nevanlinna (1982): *Stability and accuracy of time discretizations for initial value problems*. Numer. Math., vol. 40, pp. 245-296. [IV.4], [V.2], [V.4]
- R. Jeltsch, see also G. Dahlquist & R. Jeltsch.
- M.S. Kamel, see W.H. Enright & M.S. Kamel.
- L. Kantorovich & G. Akilov (1959): *Functional Analysis in Normed Spaces*. Fizmatgiz, Moscow (German translation: Academic-Verlag, Berlin, 1964). [VI.3]
- P. Kaps (1977): *Modifizierte Rosenbrockmethoden der Ordnungen 4,5 und 6 zur numerischen Integration steifer Differentialgleichungen*. Dissertation, Univ. Innsbruck. [IV.7]
- P. Kaps & A. Ostermann (1989): *Rosenbrock methods using few LU-decompositions*. IMA J. Numer. Anal., vol. 9, pp. 15-27. [IV.7]
- P. Kaps & A. Ostermann (1990): *$L(\alpha)$ -stable variable order Rosenbrock-methods*. in: K. Strehmel, ed., *Numerical treatment of differential equations*, Teubner Texte zur Mathematik, Band 121, p. 80-91. [IV.7]
- P. Kaps, S.W.H. Poon & T.D. Bui (1985): *Rosenbrock methods for stiff ODEs: a comparison of Richardson extrapolation and embedding technique*. Computing, vol. 34, pp. 17-40. [IV.7]
- P. Kaps & P. Rentrop (1979): *Generalized Runge-Kutta methods of order four with stepsize control for stiff ordinary differential equations*. Numer. Math., vol. 33, pp. 55-68. [IV.7]
- P. Kaps & G. Wanner (1981): *A study of Rosenbrock-type methods of high order*. Numer. Math., vol. 38, pp. 279-298. [IV.7]
- O.A. Karakashian, see L.A. Bales, O.A. Karakashian & S.M. Serbin.
- T. Kato (1960): *Estimation of iterated matrices, with application to the von Neumann condition*. Numer. Math., vol. 2, pp. 22-29. [V.7]
- T. Kato (1966): *Perturbation Theory for Linear Operators*. Grundlehren der math. Wissenschaften, Bd. 132, Springer Verlag, Berlin. [V.7]
- T. Kato, see H. Fujita & T. Kato.
- S.L. Keeling (1989): *On implicit Runge-Kutta methods with a stability function having distinct real poles*. BIT, vol. 29, pp. 91-109. [IV.4]
- M.D. Kirszbraun (1934): *Ueber die zusammenziehenden und Lipschitzschen Transformationen*. Fund. Math., vol. 23, pp. 77-108. [IV.12]
- R.W. Klopfenstein, see R.L. Crane & R.W. Klopfenstein.
- A.K. Kong, see R.D. Skeel & A.K. Kong.
- J.F.B.M. Kraaijevanger (1985): *B-convergence of the implicit midpoint rule and the trapezoidal rule*. BIT, vol. 25, pp. 652-666. [IV.15]

- J.F.B.M. Kraaijevanger (1986): *Absolute monotonicity of polynomials occurring in the numerical solution of initial value problems*. Numer. Math., vol. 48, pp. 303-322. [IV.11]
- J.F.B.M. Kraaijevanger (1991): *A characterization of Lyapunov diagonal stability using Hadamard products*. Linear Alg. Appl., vol. 151, pp. 245-254. [IV.14]
- J.F.B.M. Kraaijevanger & J. Schneid (1991): *On the unique solvability of the Runge-Kutta equations*. Numer. Math., vol. 59, pp. 129-157. [IV.14], [IV.15]
- J.F.B.M. Kraaijevanger, see also K.Dekker, J.F.B.M. Kraaijevanger & J. Schneid; J.A. van de Griend & J.F.B.M. Kraaijevanger; M.Z. Liu & J.F.B.M. Kraaijevanger.
- H.O. Kreiss (1962): *Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren*. BIT, vol. 2, pp. 153-181. [V.7]
- F.T. Krogh (1966): *Predictor-Corrector methods of high order with improved stability characteristics*. J. Assoc. Comput. Mach., vol. 13, pp. 374-385. [V.1]
- L. Kronecker (1874): *Über Schaaren von quadratischen und bilinearen Formen*. Akad. der Wiss. Berlin 19. Jan. 1874, Werke vol. I, pp. 351-413. [VII.1]
- L. Kronecker (1890): *Algebraische Reduction der Schaaren bilinearer Formen*. Akad. der Wiss. Berlin 27. Nov. 1890, Werke vol. III², pp. 141-155. [VII.1]
- V.I. Krylov (1959): *Priblizhennoe Vychislenie Integralov*. Goz. Izd. Fiz.-Mat. Lit., Moscow. English translation: *Approximate calculation of integrals*. Macmillan, New York, 1962. [V.3]
- P. Kunkel & V. Mehrmann (1995): *Canonical forms for linear differential-algebraic equations with variable coefficients*. J. Comp. Appl. Math., vol. 56, pp. 225-251. [VII.1]
- P. Kunkel & V. Mehrmann (1996): *Regular solutions of nonlinear differential-algebraic equations and their numerical determination*. Preprint, TU Chemnitz-Zwickau. [VII.2]
- M.A. Kurdi (1974): *Stable high order methods for time discretization of stiff differential equations*. Thesis, Univ. of California. [IV.6]
- A. Kværnø (1990): *Runge-Kutta methods applied to fully implicit differential-algebraic equations of index 1*. Math. Comp., vol. 54, pp. 583-625. [VII.5]
- B. Lago, see A. Guillou & B. Lago.
- J.L. Lagrange (1776): *Sur l'usage des fractions continues dans le calcul intégral*. Nouv. Mém. de l'Acad. royale du Sc. et Belles-Lettres de Berlin, Oeuvres Tome quatrième, pp. 301-332. [IV.3]
- J.L. Lagrange (1788): *Mécanique analytique*. Paris, chez la Veuve Desaint, Libraire, MD-CCLXXXVIII, avec approbation et privilège du Roi. Oeuvres vol. 11 et 12. [IV.1]
- S. Lang (1962): *Introduction to differentiable manifolds*. John Wiley 1962; third and enlarged edition: *Differential and Riemannian manifolds*. Graduate Texts in Mathematics, Springer 1995. [VII.1]
- J.D. Lawson (1967): *Generalized Runge-Kutta processes for stable systems with large Lipschitz constants*. SIAM J. Numer. Anal., vol. 4, pp. 372-380. [IV.9]
- V.I. Lebedev (1989): *Explicit difference schemes with time-variable steps for solving stiff systems of Equations*. Sov. J. Numer. Anal. Math. Modelling 1989, vol. 4, N2, pp. 111-135. [IV.2]
- V.I. Lebedev (1994): *How to solve stiff systems of differential equations by explicit methods*. In: *Numerical methods and applications*, ed. by G.I. Marchuk, pp. 45-80, CRC Press 1994. [IV.2]

- V.I. Lebedev (1995): *Extremal polynomials with restrictions and optimal algorithms*. Manuscript, Russian Academy of Science, Moscow. [IV.2]
- V.I. Lebedev & S.I. Finogenov (1976): *On the utilization of ordered Tchebychef parameters in iterative methods*. Zh. Vychisl. Mat. Mat Fiziki vol. 16, Nr. 4 pp. 895-910, (in Russian). [IV.2]
- V.I. Lebedev & A.A. Medovikov (1994): *Explicit methods of second order for the solution of stiff systems of ordinary differential equations* (russian). Manuscript, Russian Academy of Science, Moscow. [IV.2]
- B. van Leer, see P. Sonneveld & B. van Leer.
- B. Leimkuhler, see C. Führer & B. Leimkuhler; C.W. Gear, G.K. Gupta & B. Leimkuhler.
- B.J. Leimkuhler & R.D. Skeel (1994): *Symplectic numerical integrators in constrained Hamiltonian systems*. J. Comput. Phys., vol. 112, pp. 117-125. [VII.8]
- M.-N. Le Roux (1980): *Méthodes multipas pour des équations paraboliques non linéaires*. Numer. Math., vol. 35, pp. 143-162. [V.8]
- R.J. LeVeque & L.N. Trefethen (1984): *On the resolvent condition in the Kreiss matrix theorem*. BIT, vol. 24, pp. 584-591. [V.7]
- R. LeVeque, see also G. Dahlquist, H. Mingyou & R. LeVeque.
- H. Lewy, see R. Courant, K. Friedrichs & H. Lewy.
- I. Lie (1990): *The stability function for multistep collocation methods*. Numer. Math., vol. 57, pp. 779-787. [V.3]
- I. Lie & S.P. Nørsett (1989): *Superconvergence for multistep collocation*. Math. of Comput., vol. 52, pp. 65-79. [V.3]
- B. Lindberg (1971): *On smoothing and extrapolation for the trapezoidal rule*. BIT, vol. 11, pp. 29-52. [IV.9]
- B. Lindberg (1972): *A simple interpolation algorithm for improvement of the numerical solution of a differential equation*. SIAM J. Numer. Anal., vol. 9, pp. 662-668. [VI.5]
- B. Lindberg (1974): *On a dangerous property of methods for stiff differential equations*. BIT, vol. 14, pp. 430-436. [IV.3]
- B. Lindberg, see also W.H. Enright, T.E. Hull & B. Lindberg.
- W. Liniger (1956): *Zur Stabilität der numerischen Integrationsmethoden für Differentialgleichungen*. Thèse, Université de Lausanne, 95 p. [V.6]
- W. Liniger & R.A. Willoughby (1970): *Efficient integration methods for stiff systems of ordinary differential equations*. SIAM J. Numer. Anal., vol. 7, pp. 47-66. [IV.8]
- W. Liniger, see also O. Nevanlinna & W. Liniger; F. Odeh & W. Liniger.
- M.Z. Liu, K. Dekker & M.N. Spijker (1987): *Suitability of Runge-Kutta methods*. J. Comp. Appl. Math., vol. 91, pp. 53-63. [IV.14]
- M.Z. Liu & J.F.B.M. Kraaijevanger (1988): *On the solvability of the systems of equations arising in implicit Runge-Kutta methods*. BIT, vol. 28, pp. 825-838. [IV.14]
- C.F. Van Loan, see G.H. Golub & C.F. Van Loan.
- L. Lopez & D. Trigiante (1989): *A projection method for the numerical solution of linear systems in separable stiff differential equations*. Intern. J. Computer Math., vol. 30, pp. 191-206. [IV.10]

- P. Lötstedt (1985): *Discretization of singular perturbation problems by BDF methods*. Report No.99, Uppsala Univ., Dept. of Comp. Sci. [VI.2]
- P. Lötstedt (1985): *On the relation between singular perturbation problems and differential-algebraic equations*. Report No.100, Uppsala Univ., Dept. of Comp. Sci. [VI.2]
- P. Lötstedt & L. Petzold (1986): *Numerical solution of nonlinear differential equations with algebraic constraints I: Convergence results for backward differentiation formulas*. Math. Comput., vol 46, pp. 491-516. [VII.3]
- Ch. Lubich (1988): *Convolution quadrature and discretized operational calculus I*. Numer. Math., vol. 52, pp. 129-145. [V.7]
- Ch. Lubich (1989): *Linearly implicit extrapolation methods for differential-algebraic systems*. Numer. Math., vol. 55, pp. 197-211. [VI.6] [VII.1]
- Ch. Lubich (1989): *h^2 -extrapolation methods for differential-algebraic systems of index 2*. Impact Comput. Sc. Eng., vol. 1, pp. 260-268. [VII.7], [VII.6]
- Ch. Lubich (1991): *On the convergence of multistep methods for nonlinear stiff differential equations*. Numer. Math., vol. 58, pp. 839-853, and Erratum (Numer. Math., vol. 61, pp. 277-279) [V.7], [V.8], [VI.2]
- Ch. Lubich (1991): *Extrapolation integrators for constrained multibody systems*. Impact Comp. Sci. Eng., vol. 3, pp. 213-234. [VII.2]
- Ch. Lubich (1991): *On projected Runge-Kutta methods for differential-algebraic equations*. BIT, vol. 31, pp. 545-550. [VII.5]
- Ch. Lubich (1993): *Integration of stiff mechanical systems by Runge-Kutta methods*. ZAMP, vol. 44, pp. 1022-1053. [VII.7]
- Ch. Lubich, see also E. Hairer, G. Bader & Ch. Lubich; E. Hairer & Ch. Lubich; E. Hairer, Ch. Lubich & M. Roche.
- Ch. Lubich, U. Nowak, U. Pöhle & Ch. Engstler (1992): *MEXX – numerical software for the integration of constrained mechanical multibody systems*. Preprint SC 92-12, Konrad-Zuse-Zentrum, Berlin. [VI.7]
- Ch. Lubich & A. Ostermann (1993): *Runge-Kutta methods for parabolic equations and convolution quadrature*. Math. Comp., vol. 60, pp. 105-131. [V.8]
- Ch. Lubich & M. Roche (1990): *Rosenbrock methods for differential-algebraic systems with solution-dependent singular matrix multiplying the derivative*. Computing, vol. 43, pp. 325-342. [VI.6]
- M. Lundh, see K. Gustafsson, M. Lundh & G. Söderlind.
- F. Macsek, see W. Auzinger, R. Frank, & F. Macsek.
- D.W. Manning (1981): *A computer technique for simulating dynamic multibody systems based on dynamic formalism*. Thesis, Univ. Waterloo, Ontario. [VII.7]
- M. Marden (1966): *Geometry of polynomials*. Mathematical Surveys, American Mathematical Society, Providence, Rhode Island, 2nd edition, 243 p. [V.7]
- A.A. Markov (1890): *On a question of Mendeleiev*. Petersb. Proceedings LXII, 1-24 (Russian). [IV.2]
- J.E. Marsden, see R. Abraham, J.E. Marsden & T. Ratiu.
- R. März (1989): *Index-2 differential-algebraic equations*. Results in Mathematics, vol. 15, pp. 149-171. [VII.1]

- R. März (1990): *Higher index differential-algebraic equations: Analysis and numerical treatment*. Banach Center Publ., 24, Numer. Anal. and Math. Modelling, pp. 199-222. [VII.1], [VII.3]
- R. März, see also E. Griepentrog & R. März.
- W.S. Massey (1980): *Singular homology theory*. Graduate Texts in Mathematics 70, Springer Verlag, 265 pp. [IV.4]
- R.I. McLachlan (1995): *On the numerical integration of ordinary differential equations by symmetric composition methods*. SIAM J. Sci. Comput., vol. 16, pp. 151-168. [VII.8]
- V. Mehrmann, see also P. Kunkel & V. Mehrmann.
- M.L. Michelsen (1976): *Semi-implicit Runge-Kutta methods for stiff systems, program description and application examples*. Inst. f. Kemiteknik, Danmarks tekniske Højskole, Lyngby. [VI.4]
- K. Miller & R.N. Miller (1981): *Moving finite elements. I*. SIAM J. Numer. Anal., vol. 18, pp. 1019-1032. [VI.6]
- H. Mingyou, see G. Dahlquist, H. Mingyou & R. LeVeque.
- G.J. Minty (1962): *On a simultaneous solution of a certain system of linear inequalities*. Proc. Amer. Math. Soc., vol. 13, pp. 11-12. [IV.12]
- E.F. Mishchenko, see L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze & E.F. Mishchenko.
- J.I. Montijano (1983): *Estudio de los metodos SIRC para la resolucion numérica de ecuaciones diferenciales de tipo stiff*. Thesis, Univ. Zaragoza. [IV.14]
- E. Moore, see also S.L. Campbell & E. Moore.
- R.E. Moore, see J.W. Daniel & R.E. Moore.
- K.W. Morton, see R.D. Richtmyer & K.W. Morton.
- H.N. Mülthei (1982): *Maximale Konvergenzordnung bei der numerischen Lösung von Anfangswertproblemen mit Splines*. Numer. Math., vol. 39, pp. 449-463. [V.3]
- H.N. Mülthei (1982): *A-stabile Kollokationsverfahren mit mehrfachen Knoten*. Computing, vol. 29, pp. 51-61. [V.3]
- S. Müller, A. Prohl, R. Rannacher & S. Turek (1994): *Implicit time-discretization of the nonstationary incompressible Navier-Stokes equations*. Proc. 10th GAMM-Workshop, Kiel, W. Hackbusch & G. Wittum eds., Vieweg. [IV.3]
- A. Murua (1995): *Partitioned half-explicit Runge-Kutta methods for differential-algebraic systems of index 2*. Submitted for publication. [VII.6], [VII.7]
- C.L. Navier (1823): *Mémoire sur les lois du mouvement des fluides* (lu à l'Acad. le 18 mars 1822). Paris, Mém. de l'Acad. Royale des Sciences, Tome VI, pp. 389-440. [V.8]
- J. von Neumann (1951): *Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes*. Math. Nachrichten, vol. 4, pp. 258-281. [IV.11]
- O. Nevanlinna (1976): *On the logarithmic norms of a matrix*. Report HTKK-MAT-A94, Helsinki Univ. of Tech. [VI.3]
- O. Nevanlinna (1976): *On error bounds for G-stable methods*. BIT, vol. 16, pp. 79-84. [V.6]
- O. Nevanlinna (1977): *On the numerical integration of nonlinear initial value problems by linear multistep methods*. BIT, vol. 17, pp. 58-71. [V.8]

- O. Nevanlinna (1985): *Matrix valued versions of a result of von Neumann with an application to time discretization*. J. Comput. Appl. Math., vol. 12& 13, pp. 475-489. [V.7]
- O. Nevanlinna & W. Liniger (1978): *Contractive methods for stiff differential equations, I*. BIT, vol. 18, pp. 457-474. [V.7]
- O. Nevanlinna & W. Liniger (1979): *Contractive methods for stiff differential equations, II*. BIT, vol. 19, pp. 53-72. [V.7]
- O. Nevanlinna & F. Odeh (1981): *Multiplier techniques for linear multistep methods*. Numer. Funct. Anal. Optim., vol. 3, pp. 377-423. [V.8]
- O. Nevanlinna, see also R. Jeltsch & O. Nevanlinna.
- K. Nipp & D. Stoffer (1995): *Invariant manifolds and global error estimates of numerical integration schemes applied to stiff systems of singular perturbation type – Part I: RK-methods*. Numer. Math., vol. 70, pp. 245-257. [VI.3]
- S.P. Nørsett (1974): *Multiple Padé approximations to the exponential function*. Report No. 4/74, Dept. of Math., Univ. of Trondheim, Norway. [IV.4]
- S.P. Nørsett (1974): *Semi-explicit Runge-Kutta methods*. Report No. 6/74, Dept. of Math., Univ. of Trondheim, Norway. [IV.6]
- S.P. Nørsett (1975): *Runge-Kutta methods with coefficients depending on the Jacobian*. Report No. 1/75, Dept. of Math., Univ. of Trondheim, Norway. [IV.7]
- S.P. Nørsett (1975): *C-polynomials for rational approximations to the exponential function*. Numer. Math., vol. 25, pp. 39-56. [IV.3]
- S.P. Nørsett (1976): *Runge-Kutta methods with a multiple real eigenvalue only*. BIT, vol. 16, pp. 388-393. [IV.8]
- S.P. Nørsett & G. Wanner (1979): *The real-pole sandwich for rational approximations and oscillation equations*. BIT, vol. 19, pp. 79-94. [IV.3], [IV.4]
- S.P. Nørsett & G. Wanner (1981): *Perturbed collocation and Runge-Kutta methods*. Numer. Math., vol. 38, pp. 193-208. [IV.5], [IV.13]
- S.P. Nørsett & A. Wolfbrandt (1977): *Attainable order of rational approximations to the exponential function with only real poles*. BIT, vol. 17, pp. 200-208. [IV.4]
- S.P. Nørsett & A. Wolfbrandt (1979): *Order conditions for Rosenbrock types methods*. Numer. Math., vol. 32, pp. 1-15. [IV.7]
- S.P. Nørsett, see also A. Iserles & S.P. Nørsett; I. Lie & S.P. Nørsett; G. Wanner, E. Hairer & S.P. Nørsett.
- U. Nowak, see P. Deuffhard & U. Nowak.
- R.M. Noyes, see J. Field & R.M. Noyes.
- F. Odeh & W. Liniger (1977): *Non-linear fixed-h stability of linear multistep formulae*. J. Math. Anal. Appl., vol. 61, pp. 691-712. [V.8]
- F. Odeh, see also O. Nevanlinna & F. Odeh.
- Ö. Ólafsson, see also T. Alishenas & Ö. Ólafsson.
- R.E. O'Malley (1974): *Introduction to Singular Perturbations*. Academic Press, New York. [VI.3]
- M.K. Ormrod & G.C. Andrews (1986): *Advent: a simulation program for constrained planar kinematic and dynamic systems*. Publications of the Amer. Soc. of Mech. Eng., 86-DET-97. [VII.7]

- J.M. Ortega & W.C. Rheinboldt (1970): *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, NewYork. [VI.3], [VII.3], [VII.4], [VII.8]
- A. Ostermann (1988): *Ueber die Wahl geeigneter Approximationen an die Jacobimatrix bei linear-impliziten Runge-Kutta-Verfahren*. Dissertation, Univ. Innsbruck, pp. 66. [IV.11]
- A. Ostermann (1990): *Continuous extensions of Rosenbrock-type methods*. Computing, vol. 44, pp. 59-68. [VI.4]
- A. Ostermann, see also E. Hairer & A. Ostermann; P. Kaps & A. Ostermann, Ch. Lubich & A. Ostermann.
- H. Padé (1892): *Sur la représentation approchée d'une fonction par des fractions rationnelles*. Première Thèse ("A Monsieur Hermite"), Ann. Ec. Norm. Sup. (3), vol. 9, Supp. 3-93, Oeuvres pp. 72-165. [IV.3]
- H. Padé (1899): *Mémoire sur les développements en fractions continues de la fonction exponentielle pouvant servir d'introduction à la théorie des fractions continues algébriques*. Ann. Ec. Norm. Sup. (3), vol. 16, pp. 395-426; Oeuvres pp. 231-262. [IV.3]
- M.A. Parseval (1799): Private communication to S.F. Lacroix. See: Lacroix, *Traité des différences et des séries*, Paris 1800, p. 377, or *Traité du calcul diff. et du calcul int.*, 2^e éd, vol. 3, p. 394, Paris 1819. Also published in Paris Mémoires présentés par divers savants à l'acad. d. sc., vol 1, (1806), p. 639.
- A. Pazy (1983): *Semigroups of Linear Operators and Applications to Partial Differential Equations*. Appl. Math. Sciences 44, Springer Verlag. [V.7]
- F. Peherstorfer (1981): *Characterization of positive quadrature formulas*. SIAM J. Math. Anal., vol. 12, pp. 935-942. [IV.13]
- O. Perron (1913): *Die Lehre von den Kettenbrüchen*. Teubner, 520 pp., 3rd ed., repr. 1977. [IV.13]
- L.R. Petzold (1982): *A description of DASSL: A Differential/Algebraic System Solver*. Proceedings of IMACS World Congress, Montreal, Canada. [VII.3], [VII.7]
- L.R. Petzold (1983): *Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations*. SIAM J. Sci. Stat. Comp., vol. 4, pp. 136-148. [IV.2]
- L.R. Petzold (1986): *Order results for implicit Runge-Kutta methods applied to differential/algebraic systems*. SIAM J. Numer. Anal., vol. 23, pp. 837-852. [VI.1], [VII.4]
- L.R. Petzold, see also U. Ascher & L.R. Petzold; K.E. Brenan, S.L. Campbell & L.R. Petzold; K.E. Brenan & L.R. Petzold; C.W. Gear, H.H. Hsu & L. Petzold; C.W. Gear & L.R. Petzold; P. Lötstedt & L. Petzold.
- Z. Picel, see B.L. Ehle & Z. Picel.
- R.J. Plemmons, see G.P. Barker, A. Berman & R.J. Plemmons.
- B. van der Pol (1926): *On "Relaxation Oscillations"*. Phil. Mag., vol. 2, pp. 978-992; reproduced in: B. van der Pol, *Selected Scientific Papers*, vol. I, North-Holland Publ. Comp. Amsterdam (1960). [VI.1]
- G. Pólya & G. Szegő (1925): *Aufgaben und Lehrsätze aus der Analysis*. Two volumes, Grundlehren Band XX, Springer Verlag, many later editions and translations. [IV.4]
- L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze & E.F. Mishchenko (1961): *The mathematical theory of optimal processes*. Fizmatgiz Moscow, english translations: Wiley 1962, Pergamon Press 1964; german translation: Oldenbourg 1964. [VII.1]
- S.W.H. Poon, see P. Kaps, S.W.H. Poon & T.D. Bui.

- A. Porter, see A. Callender, D.R. Hartree & A. Porter.
- F.A. Potra (1995): *Runge-Kutta integrators for multibody dynamics*. Mechanics of Structures and Machines, vol. 23, pp. 181-197. [VII.2]
- F.A. Potra & W.C. Rheinboldt (1990): *Differential-geometric techniques for solving differential algebraic equations*. In E.J. Haug & R.C. Deyo, eds, Real-Time Integration of Mechanical System Simulation, Springer-Verlag, Berlin, pp. 155-191. [VII.2]
- F.A. Potra & W.C. Rheinboldt (1991): *On the numerical solution of Euler-Lagrange equations*. Mech. Struct. & Mech., vol. 19(1), pp. 1-18. [VII.2]
- W.H. Press, B.P. Flannery, S.A. Teukolsky & W.T. Vetterling (1986,1989): *Numerical Recipes, the art of scientific computing (FORTRAN version)*. Cambridge University Press, 702 pp. [IV.10]
- P.J. Prince, see J.R. Dormand & P.J. Prince.
- A. Prothero & A. Robinson (1974): *On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations*. Math. of Comput., vol. 28, pp. 145-162. [IV.3], [IV.15]
- V. Puiseux (1850): *Recherches sur les fonctions algébriques*. Journal de Math. vol 15, pp. 365-480. [V.4]
- T. Ratiu, see R. Abraham, J.E. Marsden & T. Ratiu.
- P.A. Raviart, see M. Crouzeix & P.A. Raviart.
- S. Reich (1996): *Symplectic integration of constrained Hamiltonian systems by composition methods*. SIAM J. Numer. Anal., vol. 33, pp. 475-491. [VII.8]
- S. Reich (1996): *On higher-order semi-explicit symplectic partitioned Runge-Kutta methods for constrained Hamiltonian systems*. Numer. Math. [VII.8]
- S. Reich, see also U.M. Ascher, H. Chin & S. Reich.
- M. Reimer (1967): *Zur Theorie der linearen Differenzenformeln*. Math. Zeitschr., vol. 95, pp. 373-402. [V.4]
- E.Ya. Remez (1957): *General computation methods of Chebyshev approximation*. UkSSR Acad. Sci. Publ., Kiev 1957 (in Russian).
- P. Rentrop, M. Roche & G. Steinebach (1989): *The application of Rosenbrock-Wanner type methods with stepsize control in differential-algebraic equations*. Numer. Math., vol. 55, pp. 545-563. [VI.1], [VI.4]
- P. Rentrop, see also P. Kaps & P. Rentrop.
- J.D. Reymond (1989): *Implementation des méthodes Radau IIA d'ordre 7 et 9*. Diploma thesis, Univ. Geneva. [IV.10]
- W.C. Rheinboldt (1984): *Differential-algebraic systems as differential equations on manifolds*. Math. Comp., vol. 43, pp. 473-482. [VII.1]
- W.C. Rheinboldt, see J.M. Ortega & W.C. Rheinboldt; F.A. Potra & W.C. Rheinboldt.
- R.D. Richtmyer & K.W. Morton (1967): *Difference Methods for Initial-Value Problems*. Wiley-Interscience. [V.7]
- B. Riemann (1857): *Allgemeine Voraussetzungen und Hilfsmittel für die Untersuchung von Functionen unbeschränkt veränderlicher Größen*. J. f. d. r. u. angew. Math., vol. 54, pp. 101-104; Werke pp. 81-84. [V.4]

- R.E. Roberson & R. Schwertassek (1988): *Dynamics of Multibody Systems*. Springer Verlag. [VII.7]
- B.C. Robertson (1987): *Detecting stiffness with explicit Runge-Kutta formulas*. Rep. 193/87, Dept. Comp. Sci., University of Toronto. [IV.2]
- H.H. Robertson (1966): *The solution of a set of reaction rate equations*. In: J. Walsh ed.: Numer. Anal., an Introduction, Academ. Press, pp. 178-182. [IV.1], [IV.10]
- A. Robinson, see A. Prothero & A. Robinson.
- M. Roche (1988): *Rosenbrock methods for differential algebraic equations*. Numer. Math., vol. 52, pp. 45-63. [VI.4]
- M. Roche (1988): *Runge-Kutta and Rosenbrock methods for differential-algebraic equations and stiff ODEs*. Doctoral thesis, Université de Genève. [VII.5]
- M. Roche (1989): *Runge-Kutta methods for differential algebraic equations*. SIAM J. Numer. Anal., vol. 26, pp. 963-975. [VI.4]
- M. Roche, see also E. Hairer, Ch. Lubich & M. Roche; Ch. Lubich & M. Roche; P. Rentrop, M. Roche & G. Steinebach.
- H.H. Rosenbrock (1962/63): *Some general implicit processes for the numerical solution of differential equations*. Computer J., vol. 5, pp. 329-330. [IV.7]
- F. Ruamps, see M. Crouzeix & F. Ruamps.
- W.B. Rubin, see T.A. Bickart & W.B. Rubin.
- J.-P. Ryckaert, G. Ciccotti & H.J.C. Berendsen (1977): *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes*. J. Comput. Phys., vol. 23, pp. 327-341. [VII.8]
- Y. Saad (1981): *Krylov subspace methods for solving large unsymmetric linear systems*. Math. Comp., vol. 37, pp. 105-126. [IV.10]
- Y. Saad (1982): *The Lanczos biorthogonalization algorithm and other oblique projection methods for solving large unsymmetric systems*. SIAM J. Numer. Anal., vol. 19, pp. 485-506. [IV.10]
- Y. Saad, see also C.W. Gear & Y. Saad.
- I.W. Sandberg & H. Sichman (1968): *Numerical integration of systems of stiff nonlinear differential equations*. The Bell System Technical Journal, vol. 47, pp. 511-527. [IV.12]
- J.M. Sanz-Serna & M.P. Calvo (1994): *Numerical Hamiltonian Problems*. Appl. Math. and Math. Comput. 7, Chapman & Hall, 207pp. [VII.8]
- V.K. Saul'ev (1960): *Integration of parabolic type equations with the method of nets* (in Russian). Moscow, Fizmatgiz 1960. [IV.2]
- A. Sayfy, see G.J. Cooper & A. Sayfy.
- E. Schäfer (1975): *A new approach to explain the "High Irradiance Responses" of photomorphogenesis on the basis of phytochrome*. J. of Math. Biology, vol. 2, pp. 41-56. [IV.10]
- W. Schiehlen, ed. (1990): *Multibody systems handbook*. Springer Verlag, Berlin. [VII.7]
- R. Scherer (1979): *A necessary condition for B-stability*. BIT, vol. 19, pp. 111-115. [IV.3], [IV.12]
- J. Schneid (1987): *B-convergence of Lobatto IIIC formulas*. Numer. Math., vol. 51, pp. 229-235. [IV.15]

- J. Schneid, see also K.Dekker, J.F.B.M. Kraaijevanger & J. Schneid; R. Frank, J. Schneid & C.W. Ueberhuber; J.F.B.M.Kraaijevanger & J. Schneid.
- C. Schneider (1991): *ROW-methods adapted to differential-algebraic systems*. Math. Comp., vol. 56, pp. 201-213. [VI.4]
- C. Schneider (1991b): Private communication. [VI.4]
- C. Schneider (1993): *Analysis of the linearly implicit mid-point rule for differential-algebraic equations*. Electronic Transactions on Numerical Analysis, vol. 1, pp. 1-10. [VI.5]
- I.J. Schoenberg (1953): *On a Theorem of Kirszbraun and Valentine*. Amer. Math. Monthly, vol. 60, pp. 620-622. [IV.12]
- S. Scholz (1989): *Order barriers for the B-convergence of ROW methods*. Computing, vol. 41, pp. 219-235. [IV.15]
- S. Scholz, see also J.G. Verwer & S. Scholz.
- J. Schroll, see R.D. Grigorieff & J. Schroll.
- J. Schur (1918): *Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind*. J. Reine u. angew. Math., vol. 147, pp. 205-232. [V.3]
- R. Schwertassek, see R.E. Roberson & R. Schwertassek.
- S.M. Serbin, see L.A. Bales, O.A. Karakashian & S.M. Sebin.
- L.F. Shampine (1977): *Stiffness and nonstiff differential equation solvers, II: detecting stiffness with Runge-Kutta methods*. ACM TOMS, vol. 3, pp. 44-53. [IV.2]
- L.F. Shampine (1980): *Implementation of implicit formulas for the solution of ODEs*. SIAM J. Sci. Stat. Comput., vol. 1, pp. 103-118. [IV.8]
- L.F. Shampine (1981): *Evaluation of a test set for stiff ODE solvers*. ACM Trans. Math. Soft., vol. 7, pp. 409-420. [IV.10]
- L.F. Shampine (1982): *Implementation of Rosenbrock methods*. ACM Trans. Math. Soft., vol. 8, pp. 93-113. [IV.7]
- L.F. Shampine (1986): *Conservation laws and the numerical solution of ODEs*. Comp. Maths. Appls., vol. 12B., pp. 1287-1296. [VII.2]
- L.F. Shampine (1987): *Control of step size and order in extrapolation codes*. J. Comp. Appl. Math., vol. 18, pp. 3-16. [IV.9]
- L.F. Shampine & K.L. Hiebert (1977): *Detecting stiffness with the Fehlberg (4,5) formulas*. Comp. & Maths. with Appls., vol. 3, pp. 41-46. [IV.2]
- L.F. Shampine & H.A. Watts (1979): *DEPAC — design of a user oriented package of ODE solvers*. Report SAND-79-2374, Sandia Nat. Lab., Albuquerque, New Mexico. [V.5]
- H. Sichman, see I.W. Sandberg & H. Sichman.
- R.F. Sincovec, A.M. Erisman, E.L. Yip & M.A. Epton (1981): *Analysis of descriptor systems using numerical algorithms*. IEEE Trans. Aut. Control, AC-26, pp. 139-147. [VII.3]
- R.D. Skeel, see also B.J. Leimkuhler & R.D. Skeel.
- R.D. Skeel & A.K. Kong (1977): *Blended linear multistep methods*. ACM TOMS, vol. 3, pp. 326-343. [V.2], [V.3], [V.5]
- H.M. Sloate & T.A. Bickart (1973): *A-stable composite multistep methods*. J. ACM, vol. 20, pp. 7-26. [V.3]

- P.E. Sobolevskiĭ (1959): *On non-stationary equations of hydrodynamics for viscous fluid*. Doklady Akad. Nauk USSR, vol. 128, pp. 45-48. [V.8]
- G. Söderlind (1981): *On the efficient solution of nonlinear equations in numerical methods for stiff differential systems*. Report TRITA-NA-8114, Royal Inst. of Tech., Stockholm. [IV.10]
- G. Söderlind (1989): *A multi-purpose system for the numerical integration of ODEs*. Appl. Math. Comp., vol. 31, pp. 346-360. [VII.6]
- G. Söderlind & G. Dahlquist (1981): *Error propagation and stiff differential systems of singular perturbation type*. Rep. TRITA-NA-8108, Royal Inst. of Tech., Stockholm. [VI.2]
- G. Söderlind, see also C. Arévalo, C. Führer & G. Söderlind; G. Dahlquist & G. Söderlind; K. Gustafsson, M. Lundh & G. Söderlind.
- B.P. Sommeijer & J.G. Verwer (1980): *A performance evaluation of a class of Runge-Kutta-Chebyshev methods for solving semi-discrete parabolic differential equations*. Report NW91/80, Mathematisch Centrum, Amsterdam. [IV.2]
- B.P. Sommeijer (1991): *RKC, a nearly-stiff ODE solver*. Available from netlib@ornl.gov, send rkc.f from ode. [IV.2], [IV.10]
- B.P. Sommeijer, see P.J. van der Houwen & B.P. Sommeijer; J.G. Verwer, W.H. Hundsdorfer & B.P. Sommeijer.
- A. Sommerfeld (1942): *Vorlesungen über theoretische Physik*. Bd.1., Mechanik; translated from the 4th german ed.: Acad. Press. [IV.1], [VII.1]
- P. Sonneveld & B. van Leer (1985): *A minimax problem along the imaginary axis*. Nieuw Archief V. Wiskunde (4), vol. 3, pp. 19-22. [IV.2]
- G. Sottas (1984): *Dynamic adaptive selection between explicit and implicit methods when solving ODE's*. Report, Sect. de math., Univ. Genève. [IV.2]
- G. Sottas & G. Wanner (1982): *The number of positive weights of a quadrature formula*. BIT, vol. 22, pp. 339-352. [IV.13]
- J.L. Soulé, see A. Guillou & J.L. Soulé.
- M.N. Spijker (1983): *Contractivity in the numerical solution of initial value problems*. Numer. Math., vol. 42, pp. 271-290. [IV.11]
- M.N. Spijker (1985): *Feasibility and contractivity in implicit Runge-Kutta methods*. J. Comp. Appl. Math., vol. 12 et 13, pp. 563-578. [IV.14]
- M.N. Spijker (1985): *Stepsize restrictions for stability of one-step methods in the numerical solution of initial value problems*. Math. Comp., vol. 45, pp. 377-392. [IV.11]
- M.N. Spijker (1986): *The relevance of algebraic stability in implicit Runge-Kutta methods*. Teubner Texte zur Mathematik 82 (K. Strehmel, ed.), pp. 158-164. [IV.15]
- M.N. Spijker (1991): *On a conjecture by LeVeque and Trefethen related to the Kreiss matrix theorem*. BIT, vol. 31, pp. 551-555. [V.7]
- M.N. Spijker, see also M. Crouzeix, W.H. Hundsdorfer & M.N. Spijker; W.H. Hundsdorfer & M.N. Spijker; M.Z. Liu, K. Dekker & M.N. Spijker.
- I.A. Stegun, see M. Abramowitz & I.A. Stegun.
- T. Steihaug & A. Wolfbrandt (1979): *An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff differential equations*. Math. Comp., vol. 33, pp. 521-534. [IV.7]

- G. Steinebach (1995): *Order-reduction of ROW-methods for DAEs and method of lines applications*. Preprint, TH Darmstadt. [VI.4]
- G. Steinebach, see P. Rentrop, M. Roche & G. Steinebach.
- B.I. Steininger, see W.H. Hundsdorfer & B.I. Steininger.
- V. Steklov (1916): *On the approximate computation of definite integrals with the help of so-called mechanical quadrature I. Convergence of mechanical quadrature formulas*. Petrograd, Bull. Acad. Sciences, ser. VI, vol. 10, pp. 169-186 (russian). See also same Journal vol. 11 (1917), pp. 557-558 for a french explanation. [IV.13]
- H.J. Stetter (1968): *Improved absolute stability of predictor-corrector schemes*. Computing, vol. 3, pp. 286-296. [V.1]
- H.J. Stetter (1973): *Analysis of discretization methods for ordinary differential equations*. Springer, Berlin. [IV.3], [IV.9], [IV.12]
- G.W. Stewart (1972): *On the sensitivity of the eigenvalue problem $Ax = \lambda Bx$* . SIAM J. Numer. Anal., vol. 9, pp. 669-686. [VII.1]
- T.J. Stieltjes (1884): *Quelques recherches sur la Théorie des quadrature dites mécaniques*. Annales Scientif. de l'Ecole Norm. Sup., troisième série, tome I, pp. 409-426. [IV.12]
- G.G. Stokes (1845): *On the theories of the internal friction of fluids in motion, and the equilibrium and motion of elastic solids*. Cambr. Phil. Soc. Trans., vol. 8. Republished in: G.G. Stokes, Mathematical and Physical Papers, vol. 1, Cambridge 1880. [V.8]
- D. Stoffer, see also K. Nipp & D. Stoffer.
- G. Strang, see A. Iserles & G. Strang.
- K. Strehmel & R. Weiner (1982): *Behandlung steifer Anfangswertprobleme gewöhnlicher Differentialgleichungen mit adaptiven Runge-Kutta Methoden*. Computing, vol. 29, pp. 153-165. [IV.11]
- K. Strehmel & R. Weiner (1987): *B-convergence results for linearly implicit one step methods*. BIT, vol. 27, pp. 264-281. [IV.11], [IV.15]
- Sun Geng (1993): *Symplectic partitioned Runge-Kutta methods*. J. Comput. Math., vol. 11, pp. 365-372. [VII.8]
- A.G. Sveshnikov, see A.N. Tikhonov, A.B. Vasil'eva & A.G. Sveshnikov.
- G. Szegő (1939): *Orthogonal Polynomials*. AMS Coll. Publ., vol. XXIII, 403pp. [IV.13]
- G. Szegő, see also G. Pólya & G. Szegő.
- E. Tadmor (1981): *The equivalence of L_2 -stability, the resolvent condition, and strict H -stability*. Lin. Alg. and its Applics., vol. 41, pp. 151-159. [V.7]
- P.G. Thomsen, see S.P. Nørsett & P.G. Thomsen.
- A.N. Tikhonov (1952): *Systems of differential equations containing small parameters in the derivatives*. Mat. Sb. (Russian), vol. 31 (73), pp. 575-586. [VI.3]
- A.N. Tikhonov, A.B. Vasil'eva & A.G. Sveshnikov (1985): *Differential Equations*. Trans. from the Russian by A.B. Sossinskij. Springer Verlag, 238pp. [VI.3]
- L.N. Trefethen, see R.J. LeVeque & L.N. Trefethen.
- D. Trigiante, see L. Lopez & D. Trigiante.
- H. Türke, see E. Hairer & H. Türke.
- C.W. Ueberhuber, see R. Frank, J. Schneid & C.W. Ueberhuber.

- R. Vanselow (1979): *Stabilitäts-und Fehleruntersuchungen bei numerischen Verfahren zur Lösung steifer nichtlinearer Anfangswertprobleme*. Diplomarbeit, Sektion Mathematik, TU-Dresden. [IV.12]
- J.M. Varah (1979): *On the efficient implementation of implicit Runge-Kutta methods*. Math. Comp., vol. 33, pp. 557-561. [IV.8]
- R.S. Varga, see G. Birkhoff & R.S. Varga.
- A.B. Vasil'eva (1963): *Asymptotic behaviour of solutions to certain problems involving nonlinear differential equations containing a small parameter multiplying the highest derivatives*. Usp. Mat. Nauk (Russian), vol. 18, pp.15-86. English translation: Russian Math. Surveys, vol.18, Nr. 3, pp. 13-84. [VI.3]
- A.B. Vasil'eva, see also A.N. Tikhonov, A.B. Vasil'eva & A.G. Sveshnikov.
- M.V. van Veldhuizen (1984): *D-stability and Kaps-Rentrop methods*. Computing vol. 32, pp. 229-237. [IV.7], [VI.4]
- L. Verlet (1967): *Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules*. Physical Review, vol. 159, pp. 98-103. [VII.8]
- J.G. Verwer (1980): *On generalized Runge-Kutta methods using an exact Jacobian at a non-step point*. ZAMM, vol. 60, pp. 263-265. [IV.7]
- J.G. Verwer (1996): *Explicit Runge-Kutta methods for parabolic partial differential equations*. To appear in Applied Numerical Mathematics. [IV.2]
- J.G. Verwer, W.H. Hundsdorfer & B.P. Sommeijer (1990): *Convergence properties of the Runge-Kutta-Chebyshev method*. Numer. Math., vol. 57, pp. 157-178. [IV.2]
- J.G. Verwer & S. Scholz (1983): *Rosenbrock methods and time-lagged Jacobian matrices*. Beiträge zur Numer. Math., vol. 11, pp. 173-183. [IV.7]
- J.G. Verwer, see also K. Dekker & J.G. Verwer.
- P.P. Wakker (1985): *Extending monotone and non-expansive mappings by optimization*. Cahiers du C.E.R.O., vol. 27, pp. 141-149. [IV.12]
- G. Wanner (1976): *A short proof on nonlinear A-stability*. BIT, vol. 16, pp. 226-227. [IV.12]
- G. Wanner (1980): *Characterization of all A-stable methods of order $2m - 4$* . BIT, vol. 20, pp. 367-374. [IV.5]
- G. Wanner, E. Hairer & S.P. Nørsett (1978): *Order stars and stability theorems*. BIT, vol. 18, pp. 475-489. [IV.4], [IV.6], [V.4]
- G. Wanner, E. Hairer & S.P. Nørsett (1978): *When I-stability implies A-stability*. BIT, vol. 18, p. 503. [IV.4]
- G. Wanner, see also E. Hairer & G. Wanner; P. Kaps & G. Wanner; S.P. Nørsett & G. Wanner; G. Sottas & G. Wanner.
- W. Wasow (1965): *Asymptotic expansions for ordinary differential equations*. Interscience, John Wiley & Sons, New York, 263pp. [VI.3]
- D.S. Watkins & R.W. HansonSmith (1983): *The numerical solution of sparably stiff systems by precise partitioning*. ACM trans. Math. Soft., vol. 9, pp. 293-301. [IV.10]
- H.A. Watts, see L.F. Shampine & H.A. Watts.
- R.A. Wehage & E.J. Haug (1982): *Generalized coordinate partitioning for dimension reduction in analysis of constrained dynamic systems*. J. Mechanical Design, vol. 104, pp. 247-255. [VII.2]

- K. Weierstrass (1868): *Zur Theorie der bilinearen und quadratischen Formen*. Akad. der Wiss. Berlin 18. Mai. 1868, Werke vol. II, pp. 19-44. [VII.1]
- R. Weiner, see K. Strehmel & R. Weiner.
- D.V. Widder (1946): *The Laplace Transform*. Princeton University Press, London. [IV.11]
- O.B. Widlund (1967): *A note on unconditionally stable linear multistep methods*. BIT, vol. 7, pp. 65-70. [IV.3], [V.1], [V.2]
- J.H. Wilkinson (1965): *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, 662 p. [IV.2]
- R.A. Williamson, see A. Iserles & R.A. Williamson.
- R.A. Willoughby (ed.) (1974): *Stiff Differential Systems*. Plenum Press, New York. [IV.1]
- R.A. Willoughby, see also W. Liniger & R.A. Willoughby.
- A. Wolfbrandt (1977): *A study of Rosenbrock processes with respect to order conditions and stiff stability*. Thesis, Chalmers Univ. of Techn., Göteborg, Sweden. [IV.4], [IV.7]
- A. Wolfbrandt, see also S.P. Nørsett & A. Wolfbrandt; T. Steihaug & A. Wolfbrandt.
- K. Wright (1970): *Some relationships between implicit Runge-Kutta, collocation and Lanczos τ methods, and their stability properties*. BIT, vol. 10, pp.217-227. [IV.3]
- J. Yen (1993): *Constrained equations of motion in multibody dynamics as ODEs on manifolds*. SIAM J. Numer. Anal., vol. 30, pp. 553-568. [VII.2]
- E.L. Yip, see R.F. Sincovec, A.M. Erisman, E.L. Yip & M.A. Epton.
- H. Yoshida (1990): *Construction of higher order symplectic integrators*. Phys. Lett. A, Vol.150, p.262-268. [VII.8]
- S. Yoshizawa, see J. Nagumo, S. Arimoto & S. Yoshizawa.
- Yuan Chzao Din (1958): *Some difference schemes of solution of first boundary problem for linear differential equations with partial derivatives*, (in Russian) Thesis cand.phys. math. Sc., Moscov MGU 1958.
- E.C. Zeeman (1972): *Differential equations for the heartbeat and nerve impulse*. Published in *Towards a theoretical biology* (Edited C.H. Waddington) Edinburgh University Press, Volume 4, pp. 8-67. Reprinted in *Catastrophe theory, Selected papers 1972-1977*, Addison-Wesley 1977, pp. 81-140. [IV.10]
- J. Zugck, see P. Deufhard, E. Hairer & J. Zugck.

Symbol Index

A	order star, 51, 285.
$A \otimes J$	tensor product, 216, 331.
B	relative order star, 59, 67, 287.
$B(p)$	simplifying assumption, 71, 363.
C	error constant, 42, 248, 262.
$C(\eta)$	simplifying assumption, 71, 363.
\mathbb{C}^+	positive half plane, 52.
\mathbb{C}^-	negative half plane, 56.
$C(\mu)$	companion matrix, 323.
$DAT, DAT2$	sets of differential algebraic trees, 410, 507.
$DAT_y, DAT2_y$	sets of differential algebraic trees, 410, 507.
$DAT_z, DAT2_z$	sets of differential algebraic trees, 410, 507.
$D(\xi)$	simplifying assumption, 71.
$D_A(\xi)$	simplifying assumption, 363.
$D_B(\xi)$	simplifying assumption, 363.
di	differentiation index, 455.
D_r	disc of radius r , 254.
$E(y)$	E -polynomial, 43, 96.
$F(t)$	elementary differential, 106, 410, 508.
$\hat{f}(\xi)$	Fourier transform, 255.
$H(p, q)$	Hamilton function, 543.
$K_q(s)$	Peano kernel, 254.
$K(Z)$	stability function for $y' = \lambda(x)y$, 185.
$LDAT, LDAT2$	sets of differential algebraic trees, 411, 509.
$LDAT_y, LDAT2_y$	sets of differential algebraic trees, 411, 509.
$LDAT_z, LDAT2_z$	sets of differential algebraic trees, 411, 509.
$\ell_i(t)$	Lagrange polynomial, 499.
$L(q, \dot{q})$	Lagrange function, 13, 463.
$L_s(x)$	Laguerre polynomial, 96.
LT_q	set of labelled trees of order q , 106.
P	projection, 494.
p_D	differentiation order, 315.
pi	perturbation index, 459.
p_I	interpolation order, 315.
$P_k(x)$	(shifted) Legendre polynomial, 78, 202.

Q	projection, 494.
$Q(\mu, \zeta)$	characteristic polynomial, 282, 291.
$R_{kj}(z)$	Padé approximation, 48.
$R(z)$	stability function, 16, 40, 41, 108, 132.
$r_j(\mu)$	coefficient of discrete resolvent, 332, 353, 385.
$r(\zeta, \mu)$	discrete resolvent, 332, 353.
S	stability domain, 16, 241.
S^{scal}	scaled stability domain, 60.
$S(Z)$	stability matrix, 353.
S_α	sector of $A(\alpha)$ -stability, 250.
$S(\mu)$	stability matrix, 290.
T	kinetic energy, 463, 531.
T	set of trees, 116.
$T_m(z)$	Chebyshev polynomial, 31.
TW	set of trees for W -methods, 115.
$T(\eta, \zeta)$	property T , 81.
U	potential energy, 463, 533.
$\ u\ _D$	norm, 218.
$\ u\ _D$	norm in product space, 216, 218.
$\ u\ _G$	norm in product space, 330.
$\ v\ _G$	inner product norm, 307, 356.
$\alpha_D(A^{-1})$	coercivity coefficient, 215.
$\alpha_0(A^{-1})$	coercivity coefficient, 215.
$\delta_D(x)$	differentiation error, 314.
$\delta_h(x)$	local error, 226, 227, 228, 323.
$\delta_I(x)$	interpolation error, 314.
$\delta_{LM}(x)$	linear multistep error, 322.
$\delta_{OL}(x)$	one-leg error, 314.
$\mu(A)$	logarithmic norm, 168.
$\mu(\zeta)$	multiplier, 343.
ν	one-sided Lipschitz constant, 180, 215, 305, 339.
ϱ	threshold factor, 176.
$\varrho(t)$	order of a tree, 410, 508.
$\varrho(\zeta)$	generating polynomial, 240.
$\sigma(\zeta)$	generating polynomial, 240.
$\varphi_B(\ell)$	error growth function, 193.
$\varphi_R(x)$	error growth function (linear problems), 169.
∇	backward difference operator, 242.

Subject Index

- A -acceptable approximations, 43.
- A -stability
 - of multistep methods, 241.
 - of one-step methods, 42f.
 - of Padé approximations, 58.
 - of rational approximations, 56f.
 - of SDIRK methods, 97.
 - via positive functions, 87.
- $A(0)$ -stable multistep methods, 250.
- A_0 -stable multistep methods, 251.
- $A(\alpha)$ -stability
 - of BDF methods, 251.
 - of blended methods, 267.
 - of Enright methods, 263.
 - of extrapolation methods, 137, 139.
 - of modified EBDf methods, 270.
 - of multistep methods, 250.
 - of multistep Radau methods, 276.
 - of RK methods, 45.
 - of second derivative BDF methods, 265.
- $A(\alpha)$ -stable multistep methods of high order, 251f.
- absolutely monotonic functions, 178.
- acceleration level, 465.
- accuracy barriers for linear multistep methods, 254f.
- Adams methods, 242f, 249, 266.
- adjoint differential equation, 462, 467.
- algebraic criterion for G -stability, 309.
- algebraic stability,
 - of general linear methods, 356f.
 - of multivalued methods, 366f.
 - of RK methods, 181f, 188, 206, 232.
- amplifier, 376f, 379.
- Andrews' squeezer mechanism, 530f.
- AN -stability,
 - of RK methods 184f, 200.
 - of general linear methods, 360.
- asymptotic expansions, 135, 428f, 433, 525f.
- asymptotic solution
 - of van der Pol's equation, 372.
- automatic stiffness detection, 21.
- backward differentiation formulas, see BDF
- backward error analysis
 - for ODEs, 555f.
 - on manifolds, 559f.
- Bader-Deuflhard method, 134f.
- Baumgarte stabilization, 470.
- B -convergence, 225.
 - of G -stable one-leg methods, 316.
 - of multistep methods, 368f.
 - of order r , 231.
 - of RK methods, 225f.
 - of trapezoidal rule, 234.
 - of variable step sizes, 230.
- BDF methods, 2-3, 239, 246, 259, 266, 280, 285, 296, 308, 477, 481, 528, 538.
- BEAM, 146, 153, 155f, 159, 300, 302.
- beam equation, 8f, 11f, 20, 38f, 46, 146.
- BECKDO, 149f, 152, 155f, 300.
- Becker-Döring model, 149f.
- Bernstein's inequality, 324.
- β -blocked multistep methods, 527.
- blended multistep methods, 266.
- boundary layer terms, 389.
- BRUSS, 148, 155f, 159f, 300, 302.
- Brusselator, 6, 19, 31, 148.
- BRUSS-2D, 151f, 157f, 160, 300.
- B -stability
 - of Radau IIA, 199.
 - of RK methods, 180f, 188, 201.
 - of Rosenbrock methods, 200.
- Burgers equation, 349f, 443f, 448.
- Cary Grant's part, 62.
- Cash's algorithm, 268.
- characteristic equation

- for general linear methods, 291.
- for linear multistep methods, 240.
- for multistep RK methods, 282.
- for predictor-corrector schemes, 244.
- characterization
 - of algebraically stable methods, 209.
 - of positive quadrature formulas, 205.
- Chebyshev method, 31f.
 - of second order, 34f.
- Chebyshev polynomial, 31f.
- chemical reactions, 3.
- Christoffel-Darboux formula, 130.
- circuits, 4, 376, 379.
- coercivity coefficient 215, 368.
- collocation methods
 - for index 2 DAE, 498.
 - multi-step, 270f.
 - one-step, 47, 78.
 - projected, 503.
 - singly implicit, 129.
- companion matrix, 323.
- comparing stability domains, 58.
- comparison
 - between Chebyshev methods, 160.
 - between extrapolation methods, 159f.
 - between IRK methods, 158f.
 - between Radau codes, 158f.
 - between Rosenbrock codes, 158f.
- composite multistep methods, 267.
- composition methods 50, 554f.
- consistent initial values
 - for index 1, 374, 378.
 - for index 2, 456.
 - for mechanical systems, 535.
- constrained mechanical system, 464, 469f, 477, 524, 543.
- construction of IRK methods, 83.
- continued fraction representation, 50, 85.
- continued fractions related to quadrature formulas, 201f.
- continuous solution, see 'dense output'
- contractivity
 - for linear problems, 167f.
 - in general norms, 175.
 - see also '*B*-stability'
- control problems, 461f.
- convergence
 - for linear problems, 321f.
 - for nonlinear problems, 339f.
 - of *A*-stable multistep methods, 317f.
 - of BDF for index 2, 486.
 - of DAE Rosenbrock methods, 416f.
 - of half-explicit RK methods, 521.
 - of multistep methods for index 2, 489.
 - of multistep methods for SPP, 383f.
 - of RK for index 1, 380.
 - of RK for index 2 DAE, 496f, 504.
 - of RK methods for DAE, 394f.
 - of RK methods for SPP, 402.
 - of symplectic methods, 547, 549.
 - see also '*B*-convergence'
- coordinate partitioning, 476, 478f.
- counter-examples
 - for existence, 217.
 - for index definitions, 460f.
 - for stability properties, 199.
- criterion for *G*-stability, 309.
- CUSP, 147, 300, 302.
- cuspl catastrophe, 147.
- DAE, 373, 451.
 - overdetermined, 477.
- Dahlquist's first barrier, 299.
- Dahlquist's second barrier, 247, 286, 297, 299.
- Dahlquist's test equation, 16, 240.
- damped Chebyshev methods, 32f.
- Daniel-Moore conjecture, 51, 286, 294, 298, 364.
- DASSL, 481, 538, 541.
- DEABM, 5, 6.
- DEBDF, 301f.
- dense output, 576.
 - of DAE extrapolation methods, 438f.
 - of DAE Rosenbrock methods, 422.
 - of Enright methods, 263f.
 - of multistep collocation methods, 272.
 - of SDIRK4, 100.
- derivative feedback (*D*), 28.
- derivative array equations, 478.
- descriptor form, 464.
- diagonally implicit RK methods, 91f.
- difference-corrected BDF, 528.
- differential-algebraic equations, see DAE.
- differential equations
 - linear, 167, 321.
 - nonlinear, 180, 339.
 - of singular perturbation type, 371f.
 - on manifolds, 457, 474f, 544.
 - perturbed, 556.
 - quasilinear, 442, 576.
 - second order, 575.
 - stiff, 2f.

- with invariants, 472f.
- differentiation index, 455, 478.
- differentiation error, 314.
 - order, 315, 319.
- diffusion, 6.
- DIRK, 61, 91f, 208, 221.
- disc theorem, 58, 254.
- discrete resolvent, 332.
- discrete variation of constants formula, 332, 348f.
- DJ*-reducible RK methods, 187.
- dominant invariant subspace, 161.
- DOPRI5, 3, 19, 22f, 25f, 30, 143, 153f, 469, 471.
 - for mechanical system, 537.
- DOP853, 11f, 18, 20, 26, 29.
 - for mechanical system, 537.
- Dormand & Prince methods, 27.
- Dorodnitsyn's asymptotic formula, 374.
- drift-off phenomenon, 468f.
- dual order stars, 295.
- DUMKA, 34f.
- efficiency diagram, 154f, 159f, 301f, 537, 539.
- EKBWH-method, 163f.
- elastic beam, 146.
- electrical circuits, 4, 376, 379.
- elementary differentials, 106.
 - for index 1 DAE, 410.
 - for index 2 DAE, 508.
- embedded formula for RADAU5, 123.
- Enright & Kamel method, 163f.
- Enright methods, 261f, 266, 275f.
- E*-polynomial, 43, 96f.
 - for Padé approximation, 70.
- ϵ -embedding method, 374, 382, 407, 426.
- ϵ -expansions for SPP
 - for exact solution, 388.
 - for RK solution, 392f.
- equivalence
 - between stability concepts, 186, 188.
 - of *A* and *B* stability, 211.
 - of *A* and *G*-stability, 310f.
- error
 - local, 226, 228f, 405, 494.
 - global 226, 321, 328, 399, 403f.
- error bounds for one-leg methods, 314f.
- error constant, 247, 286f.
 - of rational approximations, 42, 52, 61, 67.
 - of second derivative multistep methods, 262.
 - for SDBDF methods, 265.
- error growth function, 193f, 200, 229.
 - for linear problems, 169f.
 - superexponential, 171, 194.
- error propagation, 229.
- Euler equations, 463.
- Euler's method 2, 15, 45, 58.
 - explicit, 2, 15, 556.
 - half-explicit, 519, 525.
 - implicit, 3, 45, 169, 247, 491, 557.
 - symplectic, 545, 557.
- Euler's polyhedral formula, 57.
- EULSIM, 140, 160.
- existence
 - of multistep solutions, 306f, 482.
 - of numerical RK solutions, 215f, 397, 521, 546.
- expansion of SPP solutions, 388f.
- experiments with multistep codes, 300.
- explicit
 - Adams methods, 242f.
 - Euler method, 2, 15.
 - Runge-Kutta methods, 16.
 - midpoint rule, 245, 249.
 - Nyström methods, 245.
- exponential fitting points, 56.
- extended BDF methods, 267.
- extended multistep methods, 267f.
- extrapolation methods, 18, 131.
 - for index 1 DAE, 426f.
 - for quasilinear DAE, 447.
- GBS, 18.
- E5, 145, 153f, 300f.
- first integral, 472
- Fortran codes, 565.
- Fourier transform, 148, 255.
 - fast (FFT), 149, 157.
- Gauss methods, 71, 181, 184, 198, 200, 220, 226, 504.
- Gaussian quadrature formulas, 202.
- Gear & Saad method, 161f.
- general linear methods, 290f.
 - algebraic stability of, 356f.
- generalized multistep methods, 261.
- generating polynomials, 240.
- GGL formulation of mechanical system, 465, 478.
- global error, 226.
 - expansion for SPP, 399.
 - for Prothero & Robinson problem, 328.

- of linear multistep methods, 321.
 - of one-leg methods, 322.
- Graeco-Roman transformation, 256.
- Green's function, 9.
- GRK4A, 110.
- Gronwall lemma, 460.
- G -stability,
 - of one-leg methods, 307f.
 - of BDF2 method, 308, 312.
 - of general linear methods, 356.
- half-explicit methods, 519f.
 - extrapolation methods, 525.
- multistep methods, 527.
- Runge-Kutta methods, 520.
- Hamiltonian function, 473, 543.
 - perturbed, 558.
- Hamiltonian systems, 472f.
 - constrained, 543f.
 - perturbed, 558.
- hanging rope, 13f.
- HEM5, 538.
- Hermite interpolation, 271.
- Hessenberg form, 122.
- HEX5, 538.
- hidden manifold, 454.
- high order $A(\alpha)$ -stable multistep methods, 251f.
- high oscillations, 11.
- HIHA5, method of Higham & Hall, 26f.
- HIRE5, 144f, 152f, 159f, 300f.
- HLR89, 459.
- hump, 113, 405.
- hybrid multistep methods, 267.
- hyperbolic problems, 37, 51.
- implementation
 - of extrapolation schemes, 139f.
 - of IRK methods, 118f.
 - of Rosenbrock methods, 111.
- implicit
 - Adams methods, 243.
 - Euler method, 3, 45, 169, 247, 491.
 - midpoint rule, 131, 306.
 - Milne-Simpson methods, 245, 249.
 - RK methods, 40f, 71f.
- implicit differential equations
 - $Mu' = \varphi(u)$, 103, 127, 141, 376, 378f, 408, 426.
 - $M(u)u' = \varphi(u)$, 442f, 460, 576.
 - $F(u', u) = 0$, 452, 459, 478.
- inconsistent initial values
 - for DAE Rosenbrock methods, 422f.
- index, 452f.
 - differentiation, 454f.
 - index 1, 371f, 374, 445, 455, 459, 465, 537.
 - index 2, 456, 458, 460, 464, 519, 537.
 - index 3, 456, 458, 464, 537.
 - of nilpotency, 454.
 - perturbation, 459.
- index reduction, 468f.
- inexact Jacobian, 114.
- influence of perturbations, 218, 484, 493.
- integral feedback (I), 28.
- interpolation error, 314.
 - order, 315, 319.
- invariants, 472.
- IRK(DAE), 376.
- irreducible RK methods, 187.
- I -stability, 43.
- Jeltsch-Nevanlinna theorem, 60, 289.
- kinetic energy, 8f, 463, 531.
 - of mechanical systems, 531, 541.
- Kirchhoff's law, 376.
- Kreiss matrix theorem, 323.
- Kreiss problem, 542.
- KS, 148f, 300, 302.
- Kuramoto-Sivashinsky equation, 148.
- Kuntzmann-Butcher methods, 42f, 71.
- labelled trees, 105, 411, 509.
- LADAMS, 301f, 304.
- Lagrange multipliers, 196f, 464.
- Lagrange theory, 8, 13, 463.
- Lagrange-Hamilton principle, 463.
- Laguerre polynomials 96, 129f.
- Lebedev's realization, 33.
- Legendre polynomials, 71, 78, 202.
- LIMEX, 448.
- linear problems
 - contractivity, 167f.
 - index, 452f, 455.
- linearly implicit
 - Euler method, 138f.
 - Euler for index 1 DAE, 426f.
 - Euler for quasilinear DAE, 448.
 - midpoint rule, 134f, 441.
 - RK method, 102.
- Lipschitz constant, 23.

- one-sided, 180.
- Lobatto IIIA methods, 42f, 75f, 185, 211, 222, 226, 504.
- Lobatto IIIA-IIIB pair, 549f, 563.
- Lobatto IIIB methods, 75f, 185, 211, 222, 226.
- Lobatto IIIC methods, 75f, 184, 198, 220, 223, 226, 403f, 504.
- local coordinates, 475.
- local error, 226, 228f, 485, 494.
- local state space form, 474.
- logarithmic norm 168, 390.
- LSODE, 143, 153f, 300f.
- LSODI, 481.
- L -stability, 44.
 - of SDIRK methods, 98.
- manifold, 457.
- matrix pencil, 452, 466.
- MEBDF, 303f.
- mechanical system, 463, 530f.
- METAN1, 140.
- metastability, 150.
- MEXX for mechanical system, 538.
- midpoint rule, 245, 249.
- Milne-Simpson methods, 245, 249.
- monotonically labelled trees, 105, 411, 509.
- Montaigne's ruff, 287.
- moving finite elements, 442f.
- multibody mechanisms, 530.
- multiderivative multistep methods, 282.
- multiple real-pole approximations, 67, 98f.
- multiplier, 342f.
 - and nonlinearities, 346.
 - construction of, 344f.
- multistep collocation methods, 270f.
 - as general linear method, 272.
 - G -stability of, 361.
- multistep methods, 239f.
 - β -blocked, 527.
 - for index 1, 382f.
 - for index 2, 481.
 - for quasilinear DAE, 446f.
 - of Radau type, 273.
- multistep Runge-Kutta methods, 281, 362.
- multistep twin, 306.
- Navier-Stokes equations, 351.
- non-autonomous ODE, 103, 141, 408.
- nonlinear perturbations, 172.
- number of positive weights of QF, 203f.
- numerical experiments, 143, 300, 403f, 536f.
- numerical work and poles, 283.
- Nyström methods, 245.
- ODE, see differential equations.
- ODEX, 6, 7.
- one-leg multistep methods, 305f.
 - error bounds for, 314.
- one-sided Lipschitz condition, 180f, 215, 305, 339, 356.
- one-sided Lipschitz constant, 180.
- one-step methods, 1f.
- optimal control problems, 461f, 467.
- optimal stability regions, 31f.
- order conditions
 - for DAE Rosenbrock methods, 415.
 - for index 2 DAE, 506f, 512, 523.
 - for Rosenbrock methods, 104f.
 - for SDIRK methods, 91f..
 - for second derivative multistep methods, 261.
- order of a tree, 410, 508.
- order of B -convergence, 231.
- order of a quadrature formula, 202.
- order reduction, 225.
 - for Rosenbrock methods, 236.
- order stars, 51f.
 - dual, 295.
 - for BDF2, 285.
 - for general linear methods, 290.
 - for multistep methods, 279, 284f.
 - for one-step methods, 51.
 - for Padé approximations, 53.
 - for SDIRK methods, 55, 101.
 - relative, 59, 69, 287.
- order tableau
 - for DAE extrapolation methods, 431f, 441.
- OREGO, 144, 152f, 159, 300f.
- Oregonator, 13.
- overdetermined DAE, 477.
- Padé approximations to e^z , 48f, 170.
- parabolic problems, 31f, 349f.
- Parseval identity, 255, 259.
- partitioned Rosenbrock methods, 425.
- partitioning methods, 160.
- Peano kernel, 254f.
- pendulum, 463f, 468, 474.
- perturbation index, 459.
- perturbations
 - of linear equations, 348.

- of RK solutions, 219, 398.
- perturbed asymptotic expansions, 428f, 434, 448.
- perturbed differential equation, 556.
- perturbed Hamiltonian system, 558.
- PHEM56, 538.
- PI step size control, 28.
- PLATE, 146, 152f, 300f.
- plate differential equation, 146.
- poles representing numerical work, 283.
- position level, 464.
- positive functions, 86f, 313.
- positive quadrature formulas, 183, 201, 205.
- potential energy, 8f, 463, 533.
 - of mechanical systems, 533, 541.
- preconsistency, 359.
- predictive controller, 124.
- predictor-corrector schemes, 244.
- principal root, 285.
- principal sheet, 285, 292.
- projected collocation methods, 503.
- projected Runge-Kutta methods, 502, 515f.
- projection methods, 160.
 - for DAE, 470f.
 - for ODEs with invariants, 473.
- projections (index 2), 487, 494f.
- property C , 288f.
- property T , 81.
- proportional feedback (P), 28.
- Prothero-Robinson problem, 153, 225, 328, 427.
- quasilinear differential equation, 442f, 576.
 - index 1, 445.
- Radau IA, 72, 184, 220, 226, 403f, 504.
- Radau IIA, 74, 184, 197, 220, 226, 403f, 504.
- Radau methods of multistep type, 273.
- RADAUP, 158f, 574.
- RADAU5, 4f, 46, 118f, 143, 153f, 379, 566f.
 - for mechanical system, 539, 541.
- rational approximations with real poles, 61.
- RATTLE, 548f.
- real-pole sandwich, 62.
- red-black reduction, 165.
- reduced system, 372, 374, 388.
- reducible RK methods, 187f.
- region of absolute stability, see 'stability domain'
- region of step-control stability, 26f.
- regular matrix pencil, 452, 466.
- relative order star, 59, 69, 287.
- relative separation, 161.
- resolvent (discrete), 332.
- Riemann surfaces, 279f.
- RKC, 36, 143, 153f.
- RKF4(5), 25.
- RKF5(4), 24, 26.
- ROBER, 144, 152f, 159, 300f.
- Robertson reaction, 3, 18, 144.
- RODAS, 143, 153f, 158f, 420f, 574.
- RODAS5, 143, 158f, 422.
- root locus curve, 241f.
 - for BDF methods, 246.
 - for Enright methods, 263.
 - for explicit Adams methods, 243.
 - for implicit Adams methods, 243.
 - for Milne-Simpson methods, 245.
 - for Nyström methods, 245f.
 - for SDBDF methods, 265.
- ROS4, 143.
- Rosenbrock methods, 172f.
 - comparisons, 158f.
 - contractivity, 172f.
 - for stiff problems, 102, 102f.
 - for DAE, 407f, 447.
 - order reduction, 236.
 - with inexact Jacobian, 114.
- rotation number, 204.
- Routh criterion, 89.
- Runge-Kutta methods
 - explicit, 16.
 - for index 1 problems, 375.
 - for index 2 DAE, 492f.
 - for quasilinear DAE, 446f.
 - for SPP, 392f.
 - half-explicit, 520.
 - implicit, 40f, 71f.
 - projected, 502, 515f.
- savings in linear algebra, 540.
- scaled stability domain, 60.
- Schur's criterion, 278.
- SC-stability, 24f.
 - for Dormand & Prince methods, 27.
- SDBDF, 265.
- SDIRK code, 128.
- SDIRK method, 42, 44, 91, 183, 208, 403, 504.
- SDIRK4, 100, 143, 158f.
- SECDER, 303f.
- second Dahlquist barrier, 247, 254.

- second derivative BDF methods, 265.
- second derivative multistep methods, 261.
- separably stiff problems, 161.
- SEULEX, 140, 143, 153f, 160, 575.
- SHAKE, 548.
- simplified Newton, 119f, 490.
- simplifying assumptions, 71, 80f, 183, 206f, 363.
 - for index 2 DAE, 514.
- singly diagonally implicit RK methods, 91.
- singly implicit RK methods, 128f.
- singular perturbation problems, 371f, 433.
- SIRK-methods, 128f.
- smoothing step for extrapolation, 133.
- SODEX, 140, 143, 160.
- SOLOUT, 576.
- SPP, see singular perturbation problems.
- SPRINT, 301f, 304, 481.
- S -reducible RK methods, 188.
- stability analysis
 - for Euler's method, 15.
 - for explicit RK methods, 16f.
 - for modified EBDP methods, 269.
 - for multistep methods, 240f.
 - for multistep Radau methods, 274f.
 - for multistep Runge-Kutta methods, 281f.
- stability domain, 16.
 - cross-shaped 39.
 - of Bader-Deuflhard method, 134.
 - of BDF methods, 246.
 - of modified EBDP methods, 270.
 - of Chebyshev methods, 32f.
 - of DOPRI methods, 17.
 - of Enright methods, 263.
 - of ERK methods, 17.
 - of explicit Adams methods, 243.
 - of extrapolated Euler, 139.
 - of extrapolated trapezoidal rule, 132.
 - of GBS extrapolation, 19.
 - of implicit Adams methods, 243.
 - of implicit Euler method, 246.
 - of Milne-Simpson methods, 246.
 - of multistep methods, 240f.
 - of multistep Radau methods, 276.
 - of Nyström methods, 246.
 - of Padé approximations, 52.
 - of predictor-corrector schemes, 245.
- stability function $R(z)$, 16, 84.
 - of Chebyshev methods, 32f.
 - of collocation method, 47.
 - of DIRK methods, 61.
 - of DOPRI5, 17, 26.
 - of DOP853, 18.
 - of extrapolation methods, 132f.
 - of IRK methods, 40, 84.
 - of order $\geq s$, 47.
 - of Rosenbrock methods, 108.
 - of SDIRK methods, 67, 96f.
- stability function for $y' = \lambda(x)y$
 - of IRK methods, 184f.
- stability region, see stability domain.
- stabilization
 - Baumgarte, 470.
 - by projection, 470.
 - velocity, 471f.
- stabilized explicit methods, 31f.
- stage order, 226, 369.
- starting values for Newton iteration, 120.
- state space form, 374f, 474.
- state space form method, 375f, 383.
- step size selection, 123f.
 - predictive, 124.
- step-control stability, 24f.
- stiff, 1f.
- stiff eigenvalues, 161.
- stiff eigenvectors, 161.
- stiff mechanical system, 541.
- stiff stability of multistep methods, 250.
- stiff-detest, 144.
- stiffly accurate, 227, 552.
 - RK methods, 45, 376.
 - Rosenbrock methods, 418f.
 - SDIRK methods, 92f.
- stiffness, 2, 151.
 - detection, 21.
- stopping criterion, 120.
 - for Enright & Kamel method, 164.
- STRIDE, 129.
- Sullivan, Leon, 9.
- superconvergence, 500, 554.
- superexponential, 171, 194.
- super-future point, 267.
- symplecticity, 544, 547.
- symplectic methods, 543f.
 - Euler, 545, 561.
 - Lobatto IIIA-IIIB, 550, 563.
 - second order, 548f, 558, 561f.
- tangent space parametrization, 476.
- Taylor expansion
 - for index 2 DAE, 508f.
 - of DAE Rosenbrock solution, 412f.

- of DAE solutions, 411.
 - of index 2 RK solution, 510f.
- Taylor series method, 261.
- Tchébychef, see Chebyshev.
- test problems, 144f.
- theorem of von Neumann, 168, 330.
- θ -method, 42, 50.
- threshold factor, 176, 179.
- transient phase, 2.
- transistor amplifier, 376f, 379.
- trapezoidal rule, 45, 131, 185, 234, 247, 306, 357.
- trees
 - for ODE, 92, 105.
 - for index 1 DAE, 409f.
 - for index 2 DAE, 507.
 - for W -methods, 115.
 - monotonically labelled, 105, 411, 509.
- underlying ODE, 455, 478.
- uniqueness
 - of multistep solutions, 306f, 482.
 - of RK solutions, 219, 397.
- van der Houwen & Sommeijer's approach, 35.
- van der Pol's equation, 4-5, 144, 372, 403, 406, 566.
- Vandermonde matrix, 78.
- VDPOL, 144, 153f, 159, 300f.
- velocity level, 464.
- velocity stabilization, 471.
- VODE, 301f.
- Volterra-Lotka model, 556.
- von Neumann's theorem, 168, 330.
- V -transformation, 78.
- W -methods, 114, 136.
- weak AN -stability, 360.
- weak instability, 245.
- Weierstrass-Kronecker form, 452.
- work-precision diagram, 154f, 159f, 301f, 537, 539.
- W -transformation, 77f, 183f.

**Springer Series in
Computational
Mathematics**

31

Editorial Board

R. Bank
R.L. Graham
J. Stoer
R. Varga
H. Yserentant

Ernst Hairer
Christian Lubich
Gerhard Wanner

Geometric Numerical Integration

Structure-Preserving Algorithms
for Ordinary Differential Equations

Second Edition

With 146 Figures

 Springer

Ernst Hairer
Gerhard Wanner
Section de Mathématiques
Université de Genève
2-4 rue du Lièvre, C.P. 64
CH-1211 Genève 4, Switzerland
email: Ernst.Hairer@math.unige.ch
Gerhard.Wanner@math.unige.ch

Christian Lubich
Mathematisches Institut
Universität Tübingen
Auf der Morgenstelle 10
72076 Tübingen, Germany
email: Lubich@na.uni-tuebingen.de

Library of Congress Control Number: 2005938386

Mathematics Subject Classification (2000): 65Lxx, 65P10, 70Fxx, 34Cxx

ISSN 0179-3632

ISBN-10 3-540-30663-3 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-30663-4 Springer Berlin Heidelberg New York

ISBN-10 3-540-43003-2 1st Edition Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com

© Springer-Verlag Berlin Heidelberg 2002, 2004, 2006

Printed in The Netherlands

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: by the authors and TechBooks using a Springer L^AT_EX macro package

Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper SPIN: 11592242 46/TechBooks 5 4 3 2 1 0

Preface to the First Edition

They throw geometry out the door, and it comes back through the window.

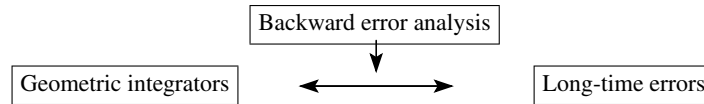
(H.G.Forder, Auckland 1973, reading new mathematics at the age of 84)

The subject of this book is numerical methods that preserve geometric properties of the flow of a differential equation: symplectic integrators for Hamiltonian systems, symmetric integrators for reversible systems, methods preserving first integrals and numerical methods on manifolds, including Lie group methods and integrators for constrained mechanical systems, and methods for problems with highly oscillatory solutions. Structure preservation – with its questions as to where, how, and what for – is the unifying theme.

In the last few decades, the theory of numerical methods for general (non-stiff and stiff) ordinary differential equations has reached a certain maturity, and excellent general-purpose codes, mainly based on Runge–Kutta methods or linear multistep methods, have become available. The motivation for developing structure-preserving algorithms for special classes of problems came independently from such different areas of research as astronomy, molecular dynamics, mechanics, theoretical physics, and numerical analysis as well as from other areas of both applied and pure mathematics. It turned out that the preservation of geometric properties of the flow not only produces an improved qualitative behaviour, but also allows for a more accurate long-time integration than with general-purpose methods.

An important shift of view-point came about by ceasing to concentrate on the numerical approximation of a single solution trajectory and instead to consider a numerical method as a *discrete dynamical system* which approximates the flow of the differential equation – and so the geometry of phase space comes back again through the window. This view allows a clear understanding of the preservation of invariants and of methods on manifolds, of symmetry and reversibility of methods, and of the symplecticity of methods and various generalizations. These subjects are presented in Chapters IV through VII of this book. Chapters I through III are of an introductory nature and present examples and numerical integrators together with important parts of the classical order theories and their recent extensions. Chapter VIII deals with questions of numerical implementations and numerical merits of the various methods.

It remains to explain the relationship between geometric properties of the numerical method and the favourable error propagation in long-time integrations. This



is done using the idea of *backward error analysis*, where the numerical one-step map is interpreted as (almost) the flow of a modified differential equation, which is constructed as an asymptotic series (Chapter IX). In this way, geometric properties of the numerical integrator translate into structure preservation on the level of the modified equations. Much insight and rigorous error estimates over long time intervals can then be obtained by combining this backward error analysis with KAM theory and related perturbation theories. This is explained in Chapters X through XII for Hamiltonian and reversible systems. The final Chapters XIII and XIV treat the numerical solution of differential equations with high-frequency oscillations and the long-time dynamics of multistep methods, respectively.

This book grew out of the lecture notes of a course given by Ernst Hairer at the University of Geneva during the academic year 1998/99. These lectures were directed at students in the third and fourth year. The reactions of students as well as of many colleagues, who obtained the notes from the Web, encouraged us to elaborate our ideas to produce the present monograph.

We want to thank all those who have helped and encouraged us to prepare this book. In particular, Martin Hairer for his valuable help in installing computers and his expertise in Latex and Postscript, Jeff Cash and Robert Chan for reading the whole text and correcting countless scientific obscurities and linguistic errors, Haruo Yoshida for making many valuable suggestions, Stéphane Cirilli for preparing the files for all the photographs, and Bernard Duzed, the irreplaceable director of the mathematics library in Geneva. We are also grateful to many friends and colleagues for reading parts of the manuscript and for valuable remarks and discussions, in particular to Assyr Abdulle, Melanie Beck, Sergio Blanes, John Butcher, Mari Paz Calvo, Begoña Cano, Philippe Chartier, David Cohen, Peter Deuffhard, Stig Faltinsen, Francesco Fassò, Martin Gander, Marlis Hochbruck, Bulent Karasözen, Wilhelm Kaup, Ben Leimkuhler, Pierre Leone, Frank Loose, Katina Lorenz, Robert McLachlan, Ander Murua, Alexander Ostermann, Truong Linh Pham, Sebastian Reich, Chus Sanz-Serna, Zaijiu Shang, Yifa Tang, Matt West, Will Wright.

We are especially grateful to Thanh-Ha Le Thi and Dr. Martin Peters from Springer-Verlag Heidelberg for assistance, in particular for their help in getting most of the original photographs from the Oberwolfach Archive and from Springer New York, and for clarifying doubts concerning the copyright.

Preface to the Second Edition

The fast development of the subject – and the fast development of the sales of the first edition of this book – has given the authors the opportunity to prepare this second edition. First of all we have corrected several misprints and minor errors which we have discovered or which have been kindly communicated to us by several readers and colleagues. We cordially thank all of them for their help and for their interest in our work. A major point of confusion has been revealed by Robert McLachlan in his book review in SIAM Reviews.

Besides many details, which have improved the presentation throughout the book, there are the following major additions and changes which make the book about 130 pages longer:

- a more prominent place of the Störmer–Verlet method in the exposition and the examples of the first chapter;
- a discussion of the Hénon–Heiles model as an example of a chaotic Hamiltonian system;
- a new Sect. IV.9 on geometric numerical linear algebra considering differential equations on Stiefel and Grassmann manifolds and dynamical low-rank approximations;
- a new improved composition method of order 10 in Sect. V.3;
- a characterization of B-series methods that conserve quadratic first integrals and a criterion for conjugate symplecticity in Sect. VI.8;
- the section on volume preservation taken from Chap. VII to Chap. VI;
- an extended and more coherent Chap. VII, renamed Non-Canonical Hamiltonian Systems, with more emphasis on the relationships between Hamiltonian systems on manifolds and Poisson systems;
- a completely reorganized and augmented Sect. VII.5 on the rigid body dynamics and Lie–Poisson systems;
- a new Sect. VII.6 on reduced Hamiltonian models of quantum dynamics and Poisson integrators for their numerical treatment;
- an improved step-size control for reversible methods in Sects. VIII.3.2 and IX.6;
- extension of Sect. IX.5 on modified equations of methods on manifolds to include constrained Hamiltonian systems and Lie–Poisson integrators;
- reorganization of Sects. IX.9 and IX.10; study of non-symplectic B-series methods that have a modified Hamiltonian, and counter-examples for symmetric methods showing linear growth in the energy error;

- a more precise discussion of integrable reversible systems with new examples in Chap. XI;
- extension of Chap. XIII on highly oscillatory problems to systems with several constant frequencies and to systems with non-constant mass matrix;
- a new Chap. XIV on oscillatory Hamiltonian systems with time- or solution-dependent high frequencies, emphasizing adiabatic transformations, adiabatic invariants, and adiabatic integrators;
- a completely rewritten Chap. XV with more emphasis on linear multistep methods for second order differential equations; a complete backward error analysis including parasitic modified differential equations; a study of the long-time stability and a rigorous explanation of the long-time near-conservation of energy and angular momentum.

Let us hope that this second revised edition will again meet good acceptance by our readers.

Geneva and Tübingen, October 2005

The Authors

Table of Contents

I.	Examples and Numerical Experiments	1
I.1	First Problems and Methods	1
I.1.1	The Lotka–Volterra Model	1
I.1.2	First Numerical Methods	3
I.1.3	The Pendulum as a Hamiltonian System	4
I.1.4	The Störmer–Verlet Scheme	7
I.2	The Kepler Problem and the Outer Solar System	8
I.2.1	Angular Momentum and Kepler’s Second Law	9
I.2.2	Exact Integration of the Kepler Problem	10
I.2.3	Numerical Integration of the Kepler Problem	12
I.2.4	The Outer Solar System	13
I.3	The Hénon–Heiles Model	15
I.4	Molecular Dynamics	18
I.5	Highly Oscillatory Problems	21
I.5.1	A Fermi–Pasta–Ulam Problem	21
I.5.2	Application of Classical Integrators	23
I.6	Exercises	24
II.	Numerical Integrators	27
II.1	Runge–Kutta and Collocation Methods	27
II.1.1	Runge–Kutta Methods	28
II.1.2	Collocation Methods	30
II.1.3	Gauss and Lobatto Collocation	34
II.1.4	Discontinuous Collocation Methods	35
II.2	Partitioned Runge–Kutta Methods	38
II.2.1	Definition and First Examples	38
II.2.2	Lobatto IIIA–IIIB Pairs	40
II.2.3	Nyström Methods	41
II.3	The Adjoint of a Method	42
II.4	Composition Methods	43
II.5	Splitting Methods	47
II.6	Exercises	50

III.	Order Conditions, Trees and B-Series	51
III.1	Runge–Kutta Order Conditions and B-Series	51
III.1.1	Derivation of the Order Conditions	51
III.1.2	B-Series	56
III.1.3	Composition of Methods	59
III.1.4	Composition of B-Series	61
III.1.5	The Butcher Group	64
III.2	Order Conditions for Partitioned Runge–Kutta Methods	66
III.2.1	Bi-Coloured Trees and P-Series	66
III.2.2	Order Conditions for Partitioned Runge–Kutta Methods	68
III.2.3	Order Conditions for Nyström Methods	69
III.3	Order Conditions for Composition Methods	71
III.3.1	Introduction	71
III.3.2	The General Case	73
III.3.3	Reduction of the Order Conditions	75
III.3.4	Order Conditions for Splitting Methods	80
III.4	The Baker–Campbell–Hausdorff Formula	83
III.4.1	Derivative of the Exponential and Its Inverse	83
III.4.2	The BCH Formula	84
III.5	Order Conditions via the BCH Formula	87
III.5.1	Calculus of Lie Derivatives	87
III.5.2	Lie Brackets and Commutativity	89
III.5.3	Splitting Methods	91
III.5.4	Composition Methods	92
III.6	Exercises	95
IV.	Conservation of First Integrals and Methods on Manifolds	97
IV.1	Examples of First Integrals	97
IV.2	Quadratic Invariants	101
IV.2.1	Runge–Kutta Methods	101
IV.2.2	Partitioned Runge–Kutta Methods	102
IV.2.3	Nyström Methods	104
IV.3	Polynomial Invariants	105
IV.3.1	The Determinant as a First Integral	105
IV.3.2	Isospectral Flows	107
IV.4	Projection Methods	109
IV.5	Numerical Methods Based on Local Coordinates	113
IV.5.1	Manifolds and the Tangent Space	114
IV.5.2	Differential Equations on Manifolds	115
IV.5.3	Numerical Integrators on Manifolds	116
IV.6	Differential Equations on Lie Groups	118
IV.7	Methods Based on the Magnus Series Expansion	121
IV.8	Lie Group Methods	123
IV.8.1	Crouch–Grossman Methods	124
IV.8.2	Munthe-Kaas Methods	125

	IV.8.3	Further Coordinate Mappings	128
IV.9		Geometric Numerical Integration Meets Geometric Numerical Linear Algebra	131
	IV.9.1	Numerical Integration on the Stiefel Manifold	131
	IV.9.2	Differential Equations on the Grassmann Manifold	135
	IV.9.3	Dynamical Low-Rank Approximation	137
IV.10		Exercises	139
V.		Symmetric Integration and Reversibility	143
V.1		Reversible Differential Equations and Maps	143
V.2		Symmetric Runge–Kutta Methods	146
	V.2.1	Collocation and Runge–Kutta Methods	146
	V.2.2	Partitioned Runge–Kutta Methods	148
V.3		Symmetric Composition Methods	149
	V.3.1	Symmetric Composition of First Order Methods	150
	V.3.2	Symmetric Composition of Symmetric Methods	154
	V.3.3	Effective Order and Processing Methods	158
V.4		Symmetric Methods on Manifolds	161
	V.4.1	Symmetric Projection	161
	V.4.2	Symmetric Methods Based on Local Coordinates	166
V.5		Energy – Momentum Methods and Discrete Gradients	171
V.6		Exercises	176
VI.		Symplectic Integration of Hamiltonian Systems	179
VI.1		Hamiltonian Systems	180
	VI.1.1	Lagrange’s Equations	180
	VI.1.2	Hamilton’s Canonical Equations	181
VI.2		Symplectic Transformations	182
VI.3		First Examples of Symplectic Integrators	187
VI.4		Symplectic Runge–Kutta Methods	191
	VI.4.1	Criterion of Symplecticity	191
	VI.4.2	Connection Between Symplectic and Symmetric Methods	194
VI.5		Generating Functions	195
	VI.5.1	Existence of Generating Functions	195
	VI.5.2	Generating Function for Symplectic Runge–Kutta Methods	198
	VI.5.3	The Hamilton–Jacobi Partial Differential Equation	200
	VI.5.4	Methods Based on Generating Functions	203
VI.6		Variational Integrators	204
	VI.6.1	Hamilton’s Principle	204
	VI.6.2	Discretization of Hamilton’s Principle	206
	VI.6.3	Symplectic Partitioned Runge–Kutta Methods Revisited	208
	VI.6.4	Noether’s Theorem	210

VI.7	Characterization of Symplectic Methods	212
VI.7.1	B-Series Methods Conserving Quadratic First Integrals	212
VI.7.2	Characterization of Symplectic P-Series (and B-Series)	217
VI.7.3	Irreducible Runge–Kutta Methods	220
VI.7.4	Characterization of Irreducible Symplectic Methods . . .	222
VI.8	Conjugate Symplecticity	222
VI.8.1	Examples and Order Conditions	223
VI.8.2	Near Conservation of Quadratic First Integrals	225
VI.9	Volume Preservation	227
VI.10	Exercises	233
VII.	Non-Canonical Hamiltonian Systems	237
VII.1	Constrained Mechanical Systems	237
VII.1.1	Introduction and Examples	237
VII.1.2	Hamiltonian Formulation	239
VII.1.3	A Symplectic First Order Method	242
VII.1.4	SHAKE and RATTLE	245
VII.1.5	The Lobatto IIIA - IIIB Pair	247
VII.1.6	Splitting Methods	252
VII.2	Poisson Systems	254
VII.2.1	Canonical Poisson Structure	254
VII.2.2	General Poisson Structures	256
VII.2.3	Hamiltonian Systems on Symplectic Submanifolds . . .	258
VII.3	The Darboux–Lie Theorem	261
VII.3.1	Commutativity of Poisson Flows and Lie Brackets . . .	261
VII.3.2	Simultaneous Linear Partial Differential Equations . . .	262
VII.3.3	Coordinate Changes and the Darboux–Lie Theorem . . .	265
VII.4	Poisson Integrators	268
VII.4.1	Poisson Maps and Symplectic Maps	268
VII.4.2	Poisson Integrators	270
VII.4.3	Integrators Based on the Darboux–Lie Theorem	272
VII.5	Rigid Body Dynamics and Lie–Poisson Systems	274
VII.5.1	History of the Euler Equations	275
VII.5.2	Hamiltonian Formulation of Rigid Body Motion	278
VII.5.3	Rigid Body Integrators	280
VII.5.4	Lie–Poisson Systems	286
VII.5.5	Lie–Poisson Reduction	289
VII.6	Reduced Models of Quantum Dynamics	293
VII.6.1	Hamiltonian Structure of the Schrödinger Equation . . .	293
VII.6.2	The Dirac–Frenkel Variational Principle	295
VII.6.3	Gaussian Wavepacket Dynamics	296
VII.6.4	A Splitting Integrator for Gaussian Wavepackets	298
VII.7	Exercises	301

VIII. Structure-Preserving Implementation	303
VIII.1 Dangers of Using Standard Step Size Control	303
VIII.2 Time Transformations	306
VIII.2.1 Symplectic Integration	306
VIII.2.2 Reversible Integration	309
VIII.3 Structure-Preserving Step Size Control	310
VIII.3.1 Proportional, Reversible Controllers	310
VIII.3.2 Integrating, Reversible Controllers	314
VIII.4 Multiple Time Stepping	316
VIII.4.1 Fast-Slow Splitting: the Impulse Method	317
VIII.4.2 Averaged Forces	319
VIII.5 Reducing Rounding Errors	322
VIII.6 Implementation of Implicit Methods	325
VIII.6.1 Starting Approximations	326
VIII.6.2 Fixed-Point Versus Newton Iteration	330
VIII.7 Exercises	335
IX. Backward Error Analysis and Structure Preservation	337
IX.1 Modified Differential Equation – Examples	337
IX.2 Modified Equations of Symmetric Methods	342
IX.3 Modified Equations of Symplectic Methods	343
IX.3.1 Existence of a Local Modified Hamiltonian	343
IX.3.2 Existence of a Global Modified Hamiltonian	344
IX.3.3 Poisson Integrators	347
IX.4 Modified Equations of Splitting Methods	348
IX.5 Modified Equations of Methods on Manifolds	350
IX.5.1 Methods on Manifolds and First Integrals	350
IX.5.2 Constrained Hamiltonian Systems	352
IX.5.3 Lie–Poisson Integrators	354
IX.6 Modified Equations for Variable Step Sizes	356
IX.7 Rigorous Estimates – Local Error	358
IX.7.1 Estimation of the Derivatives of the Numerical Solution	360
IX.7.2 Estimation of the Coefficients of the Modified Equation	362
IX.7.3 Choice of N and the Estimation of the Local Error	364
IX.8 Long-Time Energy Conservation	366
IX.9 Modified Equation in Terms of Trees	369
IX.9.1 B-Series of the Modified Equation	369
IX.9.2 Elementary Hamiltonians	373
IX.9.3 Modified Hamiltonian	375
IX.9.4 First Integrals Close to the Hamiltonian	375
IX.9.5 Energy Conservation: Examples and Counter-Examples	379
IX.10 Extension to Partitioned Systems	381
IX.10.1 P-Series of the Modified Equation	381
IX.10.2 Elementary Hamiltonians	384
IX.11 Exercises	386

X.	Hamiltonian Perturbation Theory and Symplectic Integrators	389
X.1	Completely Integrable Hamiltonian Systems	390
X.1.1	Local Integration by Quadrature	390
X.1.2	Completely Integrable Systems	393
X.1.3	Action-Angle Variables	397
X.1.4	Conditionally Periodic Flows	399
X.1.5	The Toda Lattice – an Integrable System	402
X.2	Transformations in the Perturbation Theory for Integrable Systems	404
X.2.1	The Basic Scheme of Classical Perturbation Theory . . .	405
X.2.2	Lindstedt–Poincaré Series	406
X.2.3	Kolmogorov’s Iteration	410
X.2.4	Birkhoff Normalization Near an Invariant Torus	412
X.3	Linear Error Growth and Near-Preservation of First Integrals . .	413
X.4	Near-Invariant Tori on Exponentially Long Times	417
X.4.1	Estimates of Perturbation Series	417
X.4.2	Near-Invariant Tori of Perturbed Integrable Systems . .	421
X.4.3	Near-Invariant Tori of Symplectic Integrators	422
X.5	Kolmogorov’s Theorem on Invariant Tori	423
X.5.1	Kolmogorov’s Theorem	423
X.5.2	KAM Tori under Symplectic Discretization	428
X.6	Invariant Tori of Symplectic Maps	430
X.6.1	A KAM Theorem for Symplectic Near-Identity Maps .	431
X.6.2	Invariant Tori of Symplectic Integrators	433
X.6.3	Strongly Non-Resonant Step Sizes	433
X.7	Exercises	434
XI.	Reversible Perturbation Theory and Symmetric Integrators	437
XI.1	Integrable Reversible Systems	437
XI.2	Transformations in Reversible Perturbation Theory	442
XI.2.1	The Basic Scheme of Reversible Perturbation Theory . .	443
XI.2.2	Reversible Perturbation Series	444
XI.2.3	Reversible KAM Theory	445
XI.2.4	Reversible Birkhoff-Type Normalization	447
XI.3	Linear Error Growth and Near-Preservation of First Integrals . .	448
XI.4	Invariant Tori under Reversible Discretization	451
XI.4.1	Near-Invariant Tori over Exponentially Long Times . .	451
XI.4.2	A KAM Theorem for Reversible Near-Identity Maps . .	451
XI.5	Exercises	453
XII.	Dissipatively Perturbed Hamiltonian and Reversible Systems	455
XII.1	Numerical Experiments with Van der Pol’s Equation	455
XII.2	Averaging Transformations	458
XII.2.1	The Basic Scheme of Averaging	458
XII.2.2	Perturbation Series	459

XII.3	Attractive Invariant Manifolds	460
XII.4	Weakly Attractive Invariant Tori of Perturbed Integrable Systems	464
XII.5	Weakly Attractive Invariant Tori of Numerical Integrators	465
	XII.5.1 Modified Equations of Perturbed Differential Equations	466
	XII.5.2 Symplectic Methods	467
	XII.5.3 Symmetric Methods	469
XII.6	Exercises	469
XIII.	Oscillatory Differential Equations with Constant High Frequencies .	471
XIII.1	Towards Longer Time Steps in Solving Oscillatory Equations of Motion	471
	XIII.1.1 The Störmer–Verlet Method vs. Multiple Time Scales .	472
	XIII.1.2 Gautschi’s and Deuffhard’s Trigonometric Methods . .	473
	XIII.1.3 The Impulse Method	475
	XIII.1.4 The Mollified Impulse Method	476
	XIII.1.5 Gautschi’s Method Revisited	477
	XIII.1.6 Two-Force Methods	478
XIII.2	A Nonlinear Model Problem and Numerical Phenomena	478
	XIII.2.1 Time Scales in the Fermi–Pasta–Ulam Problem	479
	XIII.2.2 Numerical Methods	481
	XIII.2.3 Accuracy Comparisons	482
	XIII.2.4 Energy Exchange between Stiff Components	483
	XIII.2.5 Near-Conservation of Total and Oscillatory Energy . . .	484
XIII.3	Principal Terms of the Modulated Fourier Expansion	486
	XIII.3.1 Decomposition of the Exact Solution	486
	XIII.3.2 Decomposition of the Numerical Solution	488
XIII.4	Accuracy and Slow Exchange	490
	XIII.4.1 Convergence Properties on Bounded Time Intervals . .	490
	XIII.4.2 Intra-Oscillatory and Oscillatory-Smooth Exchanges . .	494
XIII.5	Modulated Fourier Expansions	496
	XIII.5.1 Expansion of the Exact Solution	496
	XIII.5.2 Expansion of the Numerical Solution	498
	XIII.5.3 Expansion of the Velocity Approximation	502
XIII.6	Almost-Invariants of the Modulated Fourier Expansions	503
	XIII.6.1 The Hamiltonian of the Modulated Fourier Expansion .	503
	XIII.6.2 A Formal Invariant Close to the Oscillatory Energy . .	505
	XIII.6.3 Almost-Invariants of the Numerical Method	507
XIII.7	Long-Time Near-Conservation of Total and Oscillatory Energy .	510
XIII.8	Energy Behaviour of the Störmer–Verlet Method	513
XIII.9	Systems with Several Constant Frequencies	516
	XIII.9.1 Oscillatory Energies and Resonances	517
	XIII.9.2 Multi-Frequency Modulated Fourier Expansions	519
	XIII.9.3 Almost-Invariants of the Modulation System	521
	XIII.9.4 Long-Time Near-Conservation of Total and Oscillatory Energies	524

XIII.10	Systems with Non-Constant Mass Matrix	526
XIII.11	Exercises	529
XIV.	Oscillatory Differential Equations with Varying High Frequencies . .	531
XIV.1	Linear Systems with Time-Dependent Skew-Hermitian Matrix . .	531
XIV.1.1	Adiabatic Transformation and Adiabatic Invariants . . .	531
XIV.1.2	Adiabatic Integrators	536
XIV.2	Mechanical Systems with Time-Dependent Frequencies	539
XIV.2.1	Canonical Transformation to Adiabatic Variables	540
XIV.2.2	Adiabatic Integrators	547
XIV.2.3	Error Analysis of the Impulse Method	550
XIV.2.4	Error Analysis of the Mollified Impulse Method	554
XIV.3	Mechanical Systems with Solution-Dependent Frequencies	555
XIV.3.1	Constraining Potentials	555
XIV.3.2	Transformation to Adiabatic Variables	558
XIV.3.3	Integrators in Adiabatic Variables	563
XIV.3.4	Analysis of Multiple Time-Stepping Methods	564
XIV.4	Exercises	564
XV.	Dynamics of Multistep Methods	567
XV.1	Numerical Methods and Experiments	567
XV.1.1	Linear Multistep Methods	567
XV.1.2	Multistep Methods for Second Order Equations	569
XV.1.3	Partitioned Multistep Methods	572
XV.2	The Underlying One-Step Method	573
XV.2.1	Strictly Stable Multistep methods	573
XV.2.2	Formal Analysis for Weakly Stable Methods	575
XV.3	Backward Error Analysis	576
XV.3.1	Modified Equation for Smooth Numerical Solutions . .	576
XV.3.2	Parasitic Modified Equations	579
XV.4	Can Multistep Methods be Symplectic?	585
XV.4.1	Non-Symplecticity of the Underlying One-Step Method	585
XV.4.2	Symplecticity in the Higher-Dimensional Phase Space .	587
XV.4.3	Modified Hamiltonian of Multistep Methods	589
XV.4.4	Modified Quadratic First Integrals	591
XV.5	Long-Term Stability	592
XV.5.1	Role of Growth Parameters	592
XV.5.2	Hamiltonian of the Full Modified System	594
XV.5.3	Long-Time Bounds for Parasitic Solution Components	596
XV.6	Explanation of the Long-Time Behaviour	600
XV.6.1	Conservation of Energy and Angular Momentum	600
XV.6.2	Linear Error Growth for Integrable Systems	601
XV.7	Practical Considerations	602
XV.7.1	Numerical Instabilities and Resonances	602
XV.7.2	Extension to Variable Step Sizes	605

XV.8	Multi-Value or General Linear Methods	609
XV.8.1	Underlying One-Step Method and Backward Error Analysis	609
XV.8.2	Symplecticity and Symmetry	611
XV.8.3	Growth Parameters	614
XV.9	Exercises	615
Bibliography		617
Index		637

Chapter I.

Examples and Numerical Experiments

This chapter introduces some interesting examples of differential equations and illustrates different types of qualitative behaviour of numerical methods. We deliberately consider only very simple numerical methods of orders 1 and 2 to emphasize the qualitative aspects of the experiments. The same effects (on a different scale) occur with more sophisticated higher-order integration schemes. The experiments presented here should serve as a motivation for the theoretical and practical investigations of later chapters. The reader is encouraged to repeat the experiments or to invent similar ones.

I.1 First Problems and Methods

Numerical applications of the case of two dependent variables are not easily obtained. (A.J. Lotka 1925, p. 79)

Our first problems, the Lotka–Volterra model and the pendulum equation, are differential equations in two dimensions and show already many interesting geometric properties. Our first methods are various variants of the Euler method, the midpoint rule, and the Störmer–Verlet scheme.

I.1.1 The Lotka–Volterra Model

We start with an equation from mathematical biology which models the growth of animal species. If a real variable $u(t)$ is to represent the number of individuals of a certain species at time t , the simplest assumption about its evolution is $du/dt = u \cdot \alpha$, where α is the reproduction rate. A constant α leads to exponential growth. In the case of more species living together, the reproduction rates will also depend on the population numbers of the *other* species. For example, for two species with $u(t)$ denoting the number of predators and $v(t)$ the number of prey, a plausible assumption is made by the *Lotka–Volterra model*

$$\begin{aligned}\dot{u} &= u(v - 2) \\ \dot{v} &= v(1 - u),\end{aligned}\tag{1.1}$$

where the dots on u and v stand for differentiation with respect to time. (We have chosen the constants 2 and 1 in (1.1) arbitrarily.) A.J. Lotka (1925, Chap. VIII) used

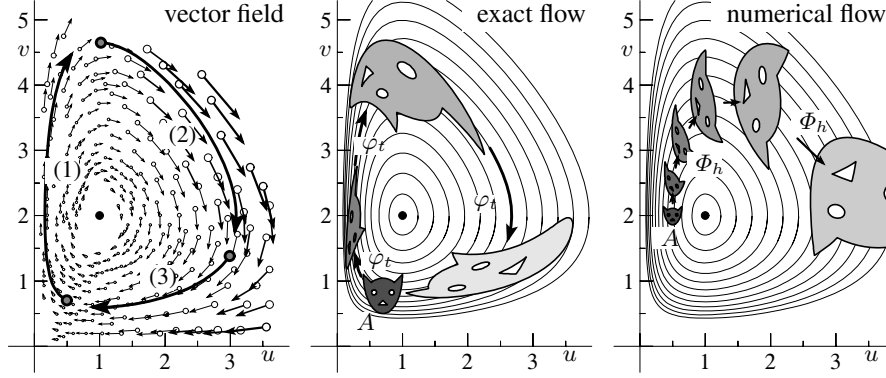


Fig. 1.1. Vector field, exact flow, and numerical flow for the Lotka–Volterra model (1.1)

this model to study parasitic invasion of insect species, and, with its help, V. Volterra (1927) explained curious fishing data from the upper Adriatic Sea following World War I.

Equations (1.1) constitute an autonomous system of differential equations. In general, we write such a system in the form

$$\dot{y} = f(y) . \quad (1.2)$$

Every y represents a point in the *phase space*, in equation (1.1) above $y = (u, v)$ is in the phase plane \mathbb{R}^2 . The vector-valued function $f(y)$ represents a *vector field* which, at any point of the phase space, prescribes the velocity (direction and speed) of the solution $y(t)$ that passes through that point (see the first picture of Fig. 1.1).

For the Lotka–Volterra model, we observe that the system cycles through three stages: (1) the prey population increases; (2) the predator population increases by feeding on the prey; (3) the predator population diminishes due to lack of food.

Flow of the System. A fundamental concept is the *flow* over time t . This is the mapping which, to any point y_0 in the phase space, associates the value $y(t)$ of the solution with initial value $y(0) = y_0$. This map, denoted by φ_t , is thus defined by

$$\varphi_t(y_0) = y(t) \quad \text{if} \quad y(0) = y_0. \quad (1.3)$$

The second picture of Fig. 1.1 shows the results of three iterations of φ_t (with $t = 1.3$) for the Lotka–Volterra problem, for a set of initial values $y_0 = (u_0, v_0)$ forming an animal-shaped set A .¹

Invariants. If we divide the two equations of (1.1) by each other, we obtain a single equation between the variables u and v . After separation of variables we get

$$0 = \frac{1-u}{u} \dot{u} - \frac{v-2}{v} \dot{v} = \frac{d}{dt} I(u, v)$$

¹ This cat came to fame through Arnold (1963).

where

$$I(u, v) = \ln u - u + 2 \ln v - v, \quad (1.4)$$

so that $I(u(t), v(t)) = \text{Const}$ for all t . We call the function I an *invariant* of the system (1.1). Every solution of (1.1) thus lies on a level curve of (1.4). Some of these curves are drawn in the pictures of Fig. 1.1. Since the level curves are closed, all solutions of (1.1) are periodic.

I.1.2 First Numerical Methods

Explicit Euler Method. The simplest of all numerical methods for the system (1.2) is the method formulated by Euler (1768),

$$y_{n+1} = y_n + hf(y_n). \quad (1.5)$$

It uses a constant step size h to compute, one after the other, approximations y_1, y_2, y_3, \dots to the values $y(h), y(2h), y(3h), \dots$ of the solution starting from a given initial value $y(0) = y_0$. The method is called the *explicit Euler method*, because the approximation y_{n+1} is computed using an explicit evaluation of f at the already known value y_n . Such a formula represents a mapping

$$\Phi_h : y_n \mapsto y_{n+1},$$

which we call the *discrete* or *numerical flow*. Some iterations of the discrete flow for the Lotka–Volterra problem (1.1) (with $h = 0.5$) are represented in the third picture of Fig. 1.1.

Implicit Euler Method. The *implicit Euler method*

$$y_{n+1} = y_n + hf(y_{n+1}), \quad (1.6)$$

is known for its all-damping stability properties. In contrast to (1.5), the approximation y_{n+1} is defined implicitly by (1.6), and the implementation requires the numerical solution of a nonlinear system of equations.

Implicit Midpoint Rule. Taking the mean of y_n and y_{n+1} in the argument of f , we get the *implicit midpoint rule*

$$y_{n+1} = y_n + hf\left(\frac{y_n + y_{n+1}}{2}\right). \quad (1.7)$$

It is a *symmetric* method, which means that the formula is left unaltered after exchanging $y_n \leftrightarrow y_{n+1}$ and $h \leftrightarrow -h$ (more on symmetric methods in Chap. V).

Symplectic Euler Methods. For *partitioned* systems

$$\begin{aligned} \dot{u} &= a(u, v) \\ \dot{v} &= b(u, v), \end{aligned} \quad (1.8)$$

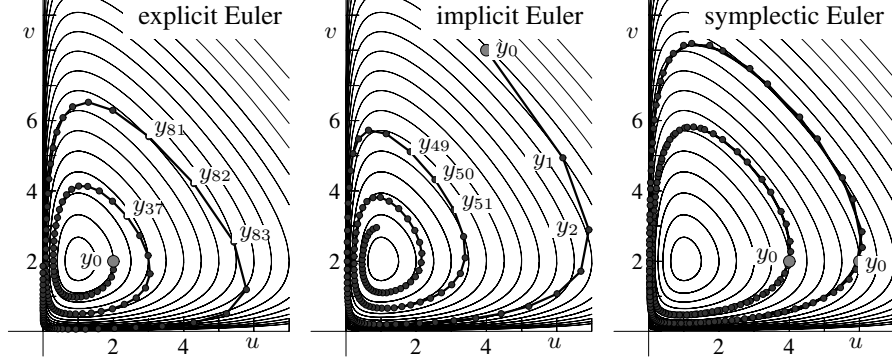


Fig. 1.2. Solutions of the Lotka–Volterra equations (1.1) (step sizes $h = 0.12$; initial values $(2, 2)$ for the explicit Euler method, $(4, 8)$ for the implicit Euler method, $(4, 2)$ and $(6, 2)$ for the symplectic Euler method)

such as the problem (1.1), we consider also *partitioned* Euler methods

$$\begin{aligned} u_{n+1} &= u_n + ha(u_n, v_{n+1}) \\ v_{n+1} &= v_n + hb(u_n, v_{n+1}), \end{aligned} \quad \text{or} \quad \begin{aligned} u_{n+1} &= u_n + ha(u_{n+1}, v_n) \\ v_{n+1} &= v_n + hb(u_{n+1}, v_n), \end{aligned} \quad (1.9)$$

which treat one variable by the implicit and the other variable by the explicit Euler method. In view of an important property of this method, discovered by de Vogelaere (1956) and to be discussed in Chap. VI, we call them *symplectic Euler methods*.

Numerical Example for the Lotka–Volterra Problem. Our first numerical experiment shows the behaviour of the various numerical methods applied to the Lotka–Volterra problem. In particular, we are interested in the preservation of the invariant I over long times. Fig. 1.2 plots the numerical approximations of the first 125 steps with the above numerical methods applied to (1.1), all with constant step sizes. We observe that the explicit and implicit Euler methods show wrong qualitative behaviour. The numerical solution either spirals outwards or inwards. The symplectic Euler method (implicit in u and explicit in v), however, gives a numerical solution that lies apparently on a closed curve as does the exact solution. Note that the curves of the numerical and exact solutions do not coincide.

I.1.3 The Pendulum as a Hamiltonian System

A great deal of attention in this book will be addressed to Hamiltonian problems, and our next examples will be of this type. These problems are of the form

$$\dot{p} = -H_q(p, q), \quad \dot{q} = H_p(p, q), \quad (1.10)$$

where the *Hamiltonian* $H(p_1, \dots, p_d, q_1, \dots, q_d)$ represents the total energy; q_i are the position coordinates and p_i the momenta for $i = 1, \dots, d$, with d the number of

degrees of freedom; H_p and H_q are the vectors of partial derivatives. One verifies easily by differentiation (see Sect. IV.1) that, along the solution curves of (1.10),

$$H(p(t), q(t)) = \text{Const}, \quad (1.11)$$

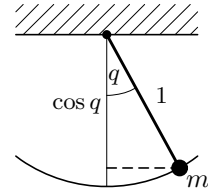
i.e., the Hamiltonian is an invariant or a *first integral*. More details about Hamiltonian systems and their derivation from Lagrangian mechanics will be given in Sect. VI.1.

Pendulum. The mathematical pendulum (mass $m = 1$, massless rod of length $\ell = 1$, gravitational acceleration $g = 1$) is a system with one degree of freedom having the Hamiltonian

$$H(p, q) = \frac{1}{2} p^2 - \cos q, \quad (1.12)$$

so that the equations of motion (1.10) become

$$\dot{p} = -\sin q, \quad \dot{q} = p. \quad (1.13)$$



Since the vector field (1.13) is 2π -periodic in q , it is natural to consider q as a variable on the circle S^1 . Hence, the phase space of points (p, q) becomes the cylinder $\mathbb{R} \times S^1$. Fig. 1.3 shows some level curves of $H(p, q)$. By (1.11), the solution curves of the problem (1.13) lie on such level curves.

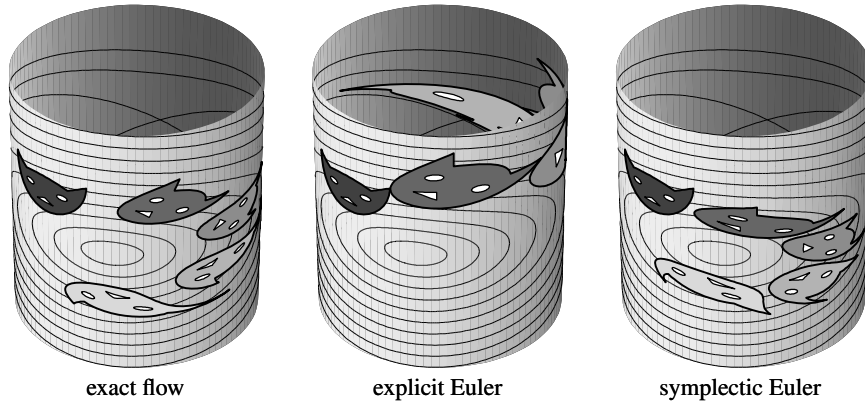


Fig. 1.3. Exact and numerical flow for the pendulum problem (1.13); step sizes $h = t = 1$

Area Preservation. Figure 1.3 (first picture) illustrates that the exact flow of a Hamiltonian system (1.10) is area preserving. This can be explained as follows: the derivative of the flow φ_t with respect to initial values (p, q) ,

$$\varphi'_t(p, q) = \frac{\partial(p(t), q(t))}{\partial(p, q)},$$

satisfies the variational equation ²

$$\dot{\varphi}'_t(p, q) = \begin{pmatrix} -H_{pq} & -H_{qq} \\ H_{pp} & H_{qp} \end{pmatrix} \varphi'_t(p, q),$$

where the second partial derivatives of H are evaluated at $\varphi_t(p, q)$. In the case of one degree of freedom ($d = 1$), a simple computation shows that

$$\frac{d}{dt} \det \varphi'_t(p, q) = \frac{d}{dt} \left(\frac{\partial p(t)}{\partial p} \frac{\partial q(t)}{\partial q} - \frac{\partial p(t)}{\partial q} \frac{\partial q(t)}{\partial p} \right) = \dots = 0.$$

Since φ_0 is the identity, this implies $\det \varphi'_t(p, q) = 1$ for all t , which means that the flow $\varphi_t(p, q)$ is an *area-preserving* mapping.

The last two pictures of Fig. 1.3 show numerical flows. The explicit Euler method is clearly seen not to preserve area but the symplectic Euler method is (this will be proved in Sect. VI.3). One of the aims of ‘geometric integration’ is the study of numerical integrators that preserve such types of qualitative behaviour of the exact flow.

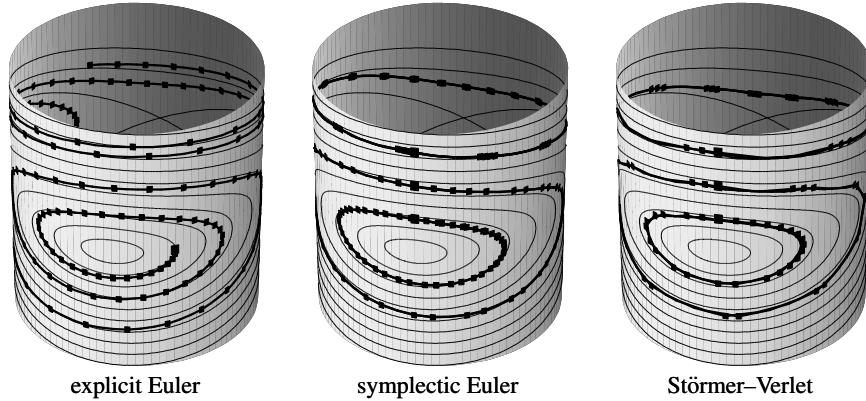


Fig. 1.4. Solutions of the pendulum problem (1.13); explicit Euler with step size $h = 0.2$, initial value $(p_0, q_0) = (0, 0.5)$; symplectic Euler with $h = 0.3$ and initial values $q_0 = 0$, $p_0 = 0.7, 1.4, 2.1$; Störmer–Verlet with $h = 0.6$

Numerical Experiment. We apply the above numerical methods to the pendulum equations (see Fig. 1.4). Similar to the computations for the Lotka–Volterra equations, we observe that the numerical solutions of the explicit Euler and of the implicit Euler method (not drawn in Fig. 1.4) spiral either outwards or inwards. The symplectic Euler method shows the correct qualitative behaviour, but destroys the left-right symmetry of the problem. The Störmer–Verlet scheme, which we discuss next, works perfectly even with doubled step size.

² As is common in the study of mechanical problems, we use *dots* for denoting time-derivatives, and we use *primes* for denoting derivatives with respect to other variables.

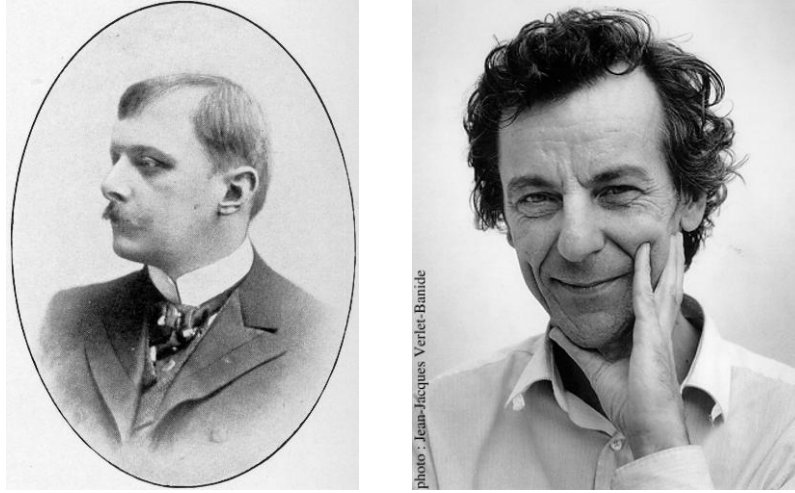


Fig. 1.5. Carl Störmer (left picture), born: 3 September 1874 in Skien (Norway), died: 13 August 1957.
Loup Verlet (right picture), born: 24 May 1931 in Paris

I.1.4 The Störmer–Verlet Scheme

The above equations (1.13) for the pendulum are of the form

$$\begin{aligned} \dot{p} &= f(q) \\ \dot{q} &= p \end{aligned} \quad \text{or} \quad \ddot{q} = f(q) \quad (1.14)$$

which is the important special case of a second order differential equation. The most natural discretization of (1.14) is

$$q_{n+1} - 2q_n + q_{n-1} = h^2 f(q_n), \quad (1.15)$$

which is just obtained by replacing the second derivative in (1.14) by the central second-order difference quotient. This basic method, or its equivalent formulation given below, is called the *Störmer method* in astronomy, the *Verlet method*³ in molecular dynamics, the *leap-frog method* in the context of partial differential equations, and it has further names in other areas (see Hairer, Lubich & Wanner (2003), p. 402). C. Störmer (1907) used higher-order variants for numerical computations concerning the aurora borealis. L. Verlet (1967) proposed this method for computations in molecular dynamics, where it has become by far the most widely used integration scheme.

Geometrically, the Störmer–Verlet method can be seen as produced by parabolas, which in the points t_n possess the right second derivative $f(q_n)$ (see Fig. 1.6

³ Irony of fate: Professor Loup Verlet, who later became interested in the history of science, discovered precisely “his” method in Newton’s *Principia* (Book I, figure for Theorem I, see Sect. I.2.1 below).

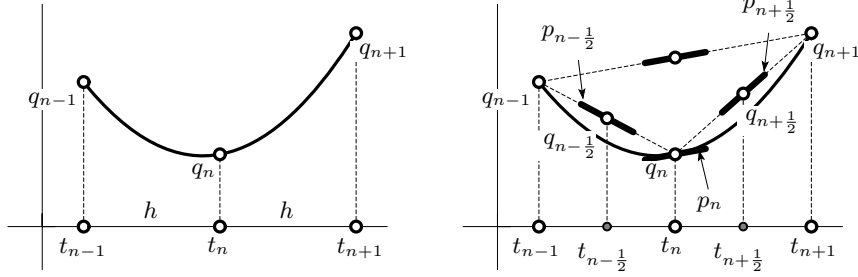


Fig. 1.6. Illustration for the Störmer-Verlet method

to the left). But we can also think of polygons, which possess the right slope in the midpoints (Fig. 1.6 to the right).

Approximations to the derivative $p = \dot{q}$ are simply obtained by

$$p_n = \frac{q_{n+1} - q_{n-1}}{2h} \quad \text{and} \quad p_{n+1/2} = \frac{q_{n+1} - q_n}{h}. \quad (1.16)$$

One-Step Formulation. The Störmer-Verlet method admits a one-step formulation which is useful for actual computations. The value q_n together with the slope p_n and the second derivative $f(q_n)$, all at t_n , uniquely determine the parabola and hence also the approximation (p_{n+1}, q_{n+1}) at t_{n+1} . Writing (1.15) as $p_{n+1/2} - p_{n-1/2} = hf(q_n)$ and using $p_{n+1/2} + p_{n-1/2} = 2p_n$, we get by elimination of either $p_{n+1/2}$ or $p_{n-1/2}$ the formulae

$$\begin{aligned} p_{n+1/2} &= p_n + \frac{h}{2} f(q_n) \\ q_{n+1} &= q_n + hp_{n+1/2} \\ p_{n+1} &= p_{n+1/2} + \frac{h}{2} f(q_{n+1}) \end{aligned} \quad (1.17)$$

which is an explicit one-step method $\Phi_h : (q_n, p_n) \mapsto (q_{n+1}, p_{n+1})$ for the corresponding first order system of (1.14). If one is not interested in the values p_n of the derivative, the first and third equations in (1.17) can be replaced by

$$p_{n+1/2} = p_{n-1/2} + hf(q_n).$$

I.2 The Kepler Problem and the Outer Solar System

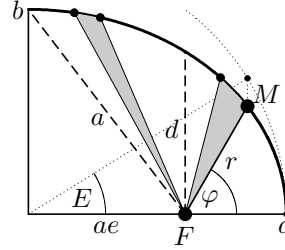
I awoke as if from sleep, a new light broke on me. (J. Kepler; quoted from J.L.E. Dreyer, *A history of astronomy*, 1906, Dover 1953, p. 391)

One of the great achievements in the history of science was the discovery of the laws of J. Kepler (1609), based on many precise measurements of the positions of Mars by Tycho Brahe and himself. The planets move in *elliptic orbits* with the sun at one of the foci (Kepler's first law)

$$r = \frac{d}{1 + e \cos \varphi} = a - ae \cos E, \quad (2.1)$$

(where a = great axis, e = eccentricity, $b = a\sqrt{1-e^2}$, $d = b\sqrt{1-e^2} = a(1-e^2)$, E = eccentric anomaly, φ = true anomaly).

Newton (*Principia* 1687) then *explained* this motion by his general law of gravitational attraction (proportional to $1/r^2$) and the relation between forces and acceleration (the “Lex II” of the *Principia*). This then opened the way for treating arbitrary celestial motions by solving differential equations.



Two-Body Problem. For computing the motion of two bodies which attract each other, we choose one of the bodies as the centre of our coordinate system; the motion will then stay in a plane (Exercise 3) and we can use two-dimensional coordinates $q = (q_1, q_2)$ for the position of the second body. Newton’s laws, with a suitable normalization, then yield the following differential equations

$$\ddot{q}_1 = -\frac{q_1}{(q_1^2 + q_2^2)^{3/2}}, \quad \ddot{q}_2 = -\frac{q_2}{(q_1^2 + q_2^2)^{3/2}}. \quad (2.2)$$

This is equivalent to a Hamiltonian system with the Hamiltonian

$$H(p_1, p_2, q_1, q_2) = \frac{1}{2} (p_1^2 + p_2^2) - \frac{1}{\sqrt{q_1^2 + q_2^2}}, \quad p_i = \dot{q}_i. \quad (2.3)$$

I.2.1 Angular Momentum and Kepler’s Second Law

The system has not only the total energy $H(p, q)$ as a first integral, but also the angular momentum

$$L(p_1, p_2, q_1, q_2) = q_1 p_2 - q_2 p_1. \quad (2.4)$$

This can be checked by differentiation and is nothing other than *Kepler’s second law*, which says that the ray FM sweeps equal areas in equal times (see the little picture at the beginning of Sect. I.2).

A beautiful *geometric* justification of this law is due to I. Newton⁴ (*Principia* (1687), Book I, figure for Theorem I). The idea is to apply the Störmer–Verlet scheme (1.15) to the equations (2.2) (see Fig. 2.1). By hypothesis, the diagonal of the parallelogram $q_{n-1}q_nq_{n+1}$, which is $(q_{n+1} - q_n) - (q_n - q_{n-1}) = q_{n+1} - 2q_n + q_{n-1} = \text{Const} \cdot f(q_n)$, points towards the sun S . Therefore, the altitudes of the triangles $q_{n-1}q_nS$ and $q_nq_{n+1}S$ are equal. Since they have the common base q_nS , they also have equal areas. Hence

$$\det(q_{n-1}, q_n - q_{n-1}) = \det(q_n, q_{n+1} - q_n)$$

and by passing to the limit $h \rightarrow 0$ we see that $\det(q, p) = \text{Const}$. This is (2.4).

⁴ We are grateful to a private communication of L. Verlet for this reference

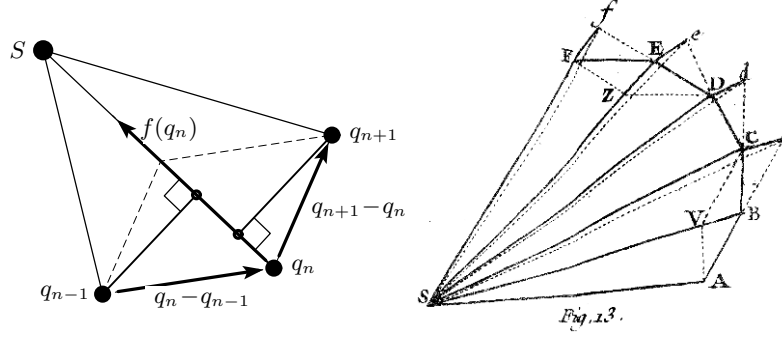


Fig. 2.1. Proof of Kepler's Second Law (left); facsimile from Newton's *Principia* (right)

We have not only an elegant proof for this invariant, but we also see that *the Störmer–Verlet scheme preserves this invariant for every $h > 0$.*

I.2.2 Exact Integration of the Kepler Problem

Pour voir présentement que cette courbe $ABC \dots$ est toujours une Section Conique, ainsi que Mr. Newton l'a supposé, *pag. 55. Coroll. I.* sans le démontrer; il y faut bien plus d'adresse: (Joh. Bernoulli 1710, p. 475)

It is now interesting, inversely to the procedure of Newton, to prove that *any* solution of (2.2) follows either an elliptic, parabolic or hyperbolic arc and to describe the solutions analytically. This was first done by Joh. Bernoulli (1710, full of sarcasm against Newton), and by Newton (1713, second edition of the *Principia*, without mentioning a word about Bernoulli).

By (2.3) and (2.4), every solution of (2.2) satisfies the two relations

$$\frac{1}{2} (\dot{q}_1^2 + \dot{q}_2^2) - \frac{1}{\sqrt{q_1^2 + q_2^2}} = H_0, \quad q_1 \dot{q}_2 - q_2 \dot{q}_1 = L_0, \quad (2.5)$$

where the constants H_0 and L_0 are determined by the initial values. Using polar coordinates $q_1 = r \cos \varphi$, $q_2 = r \sin \varphi$, this system becomes

$$\frac{1}{2} (\dot{r}^2 + r^2 \dot{\varphi}^2) - \frac{1}{r} = H_0, \quad r^2 \dot{\varphi} = L_0. \quad (2.6)$$

For its solution we consider r as a function of φ and write $\dot{r} = \frac{dr}{d\varphi} \cdot \dot{\varphi}$. The elimination of $\dot{\varphi}$ in (2.6) then yields

$$\frac{1}{2} \left(\left(\frac{dr}{d\varphi} \right)^2 + r^2 \right) \frac{L_0^2}{r^4} - \frac{1}{r} = H_0.$$

In this equation we use the substitution $r = 1/u$, $dr = -du/u^2$, which gives (with $' = d/d\varphi$)

$$\frac{1}{2} (u'^2 + u^2) - \frac{u}{L_0^2} - \frac{H_0}{L_0^2} = 0. \quad (2.7)$$

This is a “Hamiltonian” for the system

$$u'' + u = \frac{1}{d} \quad \text{i.e.,} \quad u = \frac{1}{d} + c_1 \cos \varphi + c_2 \sin \varphi = \frac{1 + e \cos(\varphi - \varphi^*)}{d} \quad (2.8)$$

where $d = L_0^2$ and the constant e becomes, from (2.7),

$$e^2 = 1 + 2H_0 L_0^2 \quad (2.9)$$

(by Exercise 7, the expression $1 + 2H_0 L_0^2$ is non-negative). This is precisely formula (2.1). The angle φ^* is determined by the initial values r_0 and φ_0 . Equation (2.1) represents an elliptic orbit with eccentricity e for $H_0 < 0$ (see Fig. 2.2, dotted line), a parabola for $H_0 = 0$, and a hyperbola for $H_0 > 0$.

Finally, we must determine the variables r and φ as functions of t . With the relation (2.8) and $r = 1/u$, the second equation of (2.6) gives

$$\frac{d^2}{(1 + e \cos(\varphi - \varphi^*))^2} d\varphi = L_0 dt \quad (2.10)$$

which, after an elementary, but not easy, integration, represents an implicit equation for $\varphi(t)$.

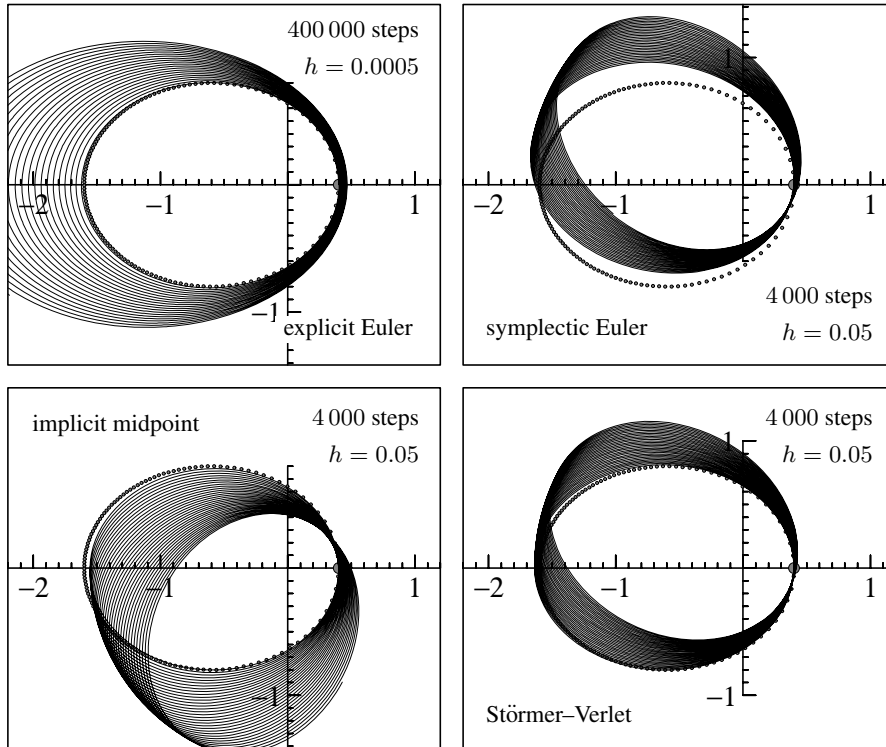


Fig. 2.2. Numerical solutions of the Kepler problem (eccentricity $e = 0.6$; in dots: exact solution)

I.2.3 Numerical Integration of the Kepler Problem

For the problem (2.2) we choose, with $0 \leq e < 1$, the initial values

$$q_1(0) = 1 - e, \quad q_2(0) = 0, \quad \dot{q}_1(0) = 0, \quad \dot{q}_2(0) = \sqrt{\frac{1+e}{1-e}}. \quad (2.11)$$

This implies that $H_0 = -1/2$, $L_0 = \sqrt{1-e^2}$, $d = 1 - e^2$ and $\varphi^* = 0$. The period of the solution is 2π (Exercise 5). Fig. 2.2 shows some numerical solutions for the eccentricity $e = 0.6$ compared to the exact solution. After our previous experience, it is no longer a surprise that the explicit Euler method spirals outwards and gives a completely wrong answer. For the other methods we take a step size 100 times larger in order to “see something”. We see that the nonsymmetric symplectic Euler method distorts the ellipse, and that all methods exhibit a *precession* effect, clockwise for Störmer–Verlet and symplectic Euler, anti-clockwise for the implicit midpoint rule. The same behaviour occurs for the exact solution of *perturbed* Kepler problems (Exercise 12) and has occupied astronomers for centuries.

Our next experiment (Fig. 2.3) studies the conservation of invariants and the global error. The main observation is that the error in the energy grows linearly for the explicit Euler method, and it remains bounded and small (no secular terms) for the symplectic Euler method. The global error, measured in the Euclidean norm, shows a quadratic growth for the explicit Euler compared to a linear growth for the symplectic Euler. As indicated in Table 2.1 the implicit midpoint rule and the Störmer–Verlet scheme behave similar to the symplectic Euler, but have a smaller

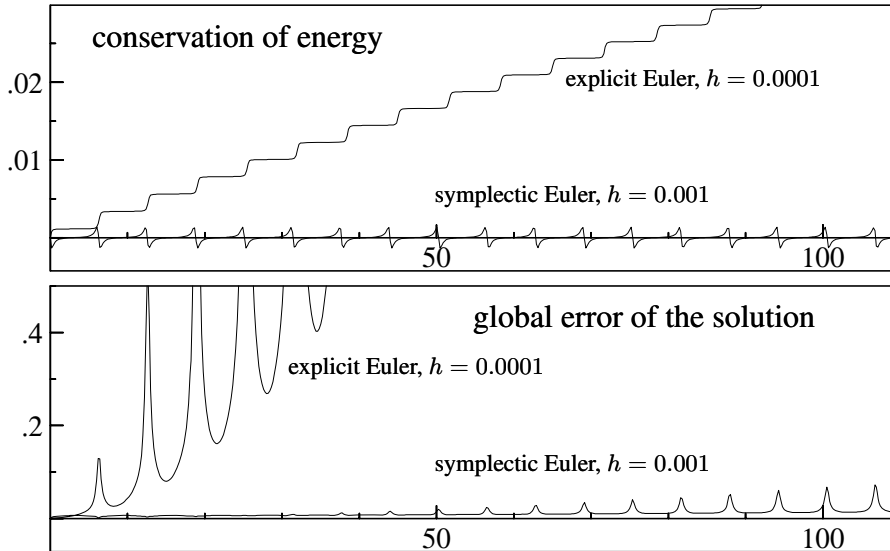


Fig. 2.3. Energy conservation and global error for the Kepler problem

Table 2.1. Qualitative long-time behaviour for the Kepler problem; t is time, h the step size

method	error in H	error in L	global error
explicit Euler	$\mathcal{O}(th)$	$\mathcal{O}(th)$	$\mathcal{O}(t^2h)$
symplectic Euler	$\mathcal{O}(h)$	0	$\mathcal{O}(th)$
implicit midpoint	$\mathcal{O}(h^2)$	0	$\mathcal{O}(th^2)$
Störmer–Verlet	$\mathcal{O}(h^2)$	0	$\mathcal{O}(th^2)$

error due to their higher order. We remark that the angular momentum $L(p, q)$ is exactly conserved by the symplectic Euler, the Störmer–Verlet, and the implicit midpoint rule.

I.2.4 The Outer Solar System

The evolution of the entire planetary system has been numerically integrated for a time span of nearly 100 million years⁵. This calculation confirms that the evolution of the solar system as a whole is chaotic, . . .
(G.J. Sussman & J. Wisdom 1992)

We next apply our methods to the system which describes the motion of the five outer planets relative to the sun. This system has been studied extensively by astronomers. The problem is a Hamiltonian system (1.10) (N -body problem) with

$$H(p, q) = \frac{1}{2} \sum_{i=0}^5 \frac{1}{m_i} p_i^T p_i - G \sum_{i=1}^5 \sum_{j=0}^{i-1} \frac{m_i m_j}{\|q_i - q_j\|}. \quad (2.12)$$

Here p and q are the supervectors composed by the vectors $p_i, q_i \in \mathbb{R}^3$ (momenta and positions), respectively. The chosen units are: masses relative to the sun, so that the sun has mass 1. We have taken

$$m_0 = 1.00000597682$$

to take account of the inner planets. Distances are in astronomical units (1 [A.U.] = 149 597 870 [km]), times in earth days, and the gravitational constant is

$$G = 2.95912208286 \cdot 10^{-4}.$$

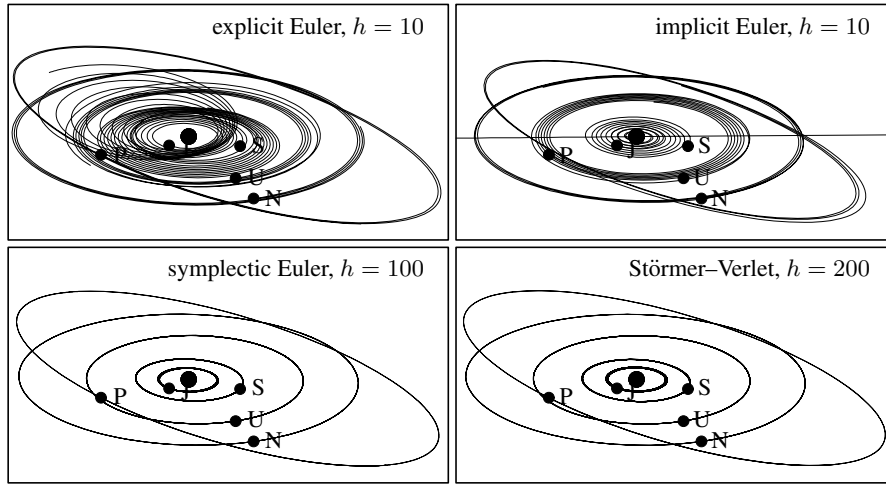
The initial values for the sun are taken as $q_0(0) = (0, 0, 0)^T$ and $\dot{q}_0(0) = (0, 0, 0)^T$. All other data (masses of the planets and the initial positions and initial velocities) are given in Table 2.2. The initial data is taken from “Ahnerts Kalender für Sternfreunde 1994”, Johann Ambrosius Barth Verlag 1993, and they correspond to September 5, 1994 at 0h00.⁶

⁵ 100 million years is not much in astronomical time scales; it just goes back to “Jurassic Park”.

⁶ We thank Alexander Ostermann, who provided us with this data.

Table 2.2. Data for the outer solar system

planet	mass	initial position	initial velocity
Jupiter	$m_1 = 0.000954786104043$	-3.5023653	0.00565429
		-3.8169847	-0.00412490
		-1.5507963	-0.00190589
Saturn	$m_2 = 0.000285583733151$	9.0755314	0.00168318
		-3.0458353	0.00483525
		-1.6483708	0.00192462
Uranus	$m_3 = 0.0000437273164546$	8.3101420	0.00354178
		-16.2901086	0.00137102
		-7.2521278	0.00055029
Neptune	$m_4 = 0.0000517759138449$	11.4707666	0.00288930
		-25.7294829	0.00114527
		-10.8169456	0.00039677
Pluto	$m_5 = 1/(1.3 \cdot 10^8)$	-15.5387357	0.00276725
		-25.2225594	-0.00170702
		-3.1902382	-0.00136504

**Fig. 2.4.** Solutions of the outer solar system

To this system we apply the explicit and implicit Euler methods with step size $h = 10$, the symplectic Euler and the Störmer-Verlet method with much larger step sizes $h = 100$ and $h = 200$, respectively, all over a time period of 200 000 days. The numerical solution (see Fig. 2.4) behaves similarly to that for the Kepler problem. With the explicit Euler method the planets have increasing energy, they spiral outwards, Jupiter approaches Saturn which leaves the plane of the two-body motion. With the implicit Euler method the planets (first Jupiter and then Saturn)

fall into the sun and are thrown far away. Both the symplectic Euler method and the Störmer–Verlet scheme show the correct behaviour. An integration over a much longer time of say several million years does not deteriorate this behaviour. Let us remark that Sussman & Wisdom (1992) have integrated the outer solar system with special geometric integrators.

I.3 The Hénon–Heiles Model

... because: (1) it is analytically simple; this makes the computation of the trajectories easy; (2) at the same time, it is sufficiently complicated to give trajectories which are far from trivial. (Hénon & Heiles 1964)

The Hénon–Heiles model was created for describing stellar motion, followed for a very long time, inside the gravitational potential $U_0(r, z)$ of a galaxy with cylindrical symmetry (Hénon & Heiles 1964). Extensive numerical experimentations should help to answer the question, if there exists, besides the known invariants H and L , a *third* invariant. Despite endless tentatives of analytical calculations during many decades, such a formula had not been found.

After a reduction of the dimension, a Hamiltonian in two degrees of freedom of the form

$$H(p, q) = \frac{1}{2}(p_1^2 + p_2^2) + U(q) \quad (3.1)$$

is obtained and the question is, if such an equation has a *second* invariant. Here, Hénon and Heiles put aside the astronomical origin of the problem and choose

$$U(q) = \frac{1}{2}(q_1^2 + q_2^2) + q_1^2 q_2 - \frac{1}{3} q_2^3 \quad (3.2)$$

(see citation). The potential U is represented in Fig. 3.1. When U approaches $\frac{1}{6}$, the level curves of U tend to an equilateral triangle, whose vertices are saddle points of U . The corresponding system

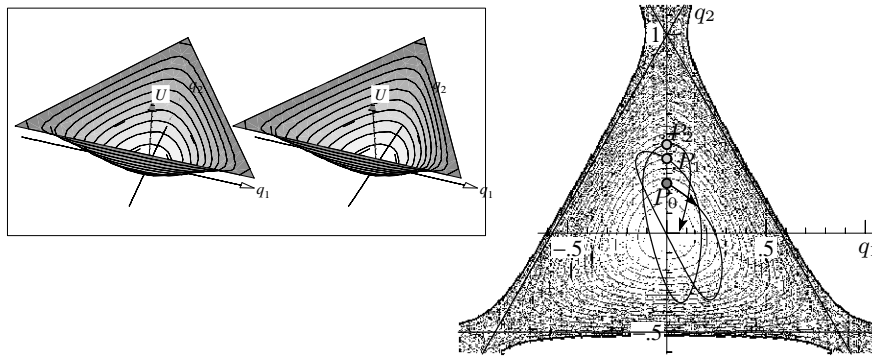


Fig. 3.1. Potential of the Hénon–Heiles Model and a solution

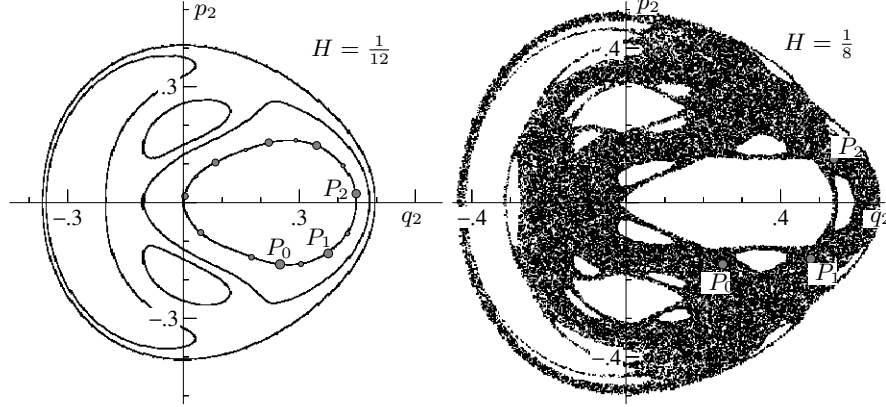


Fig. 3.2. Poincaré cuts for $q_1 = 0, p_1 > 0$ of the Hénon–Heiles Model for $H = \frac{1}{12}$ (6 orbits, left) and $H = \frac{1}{8}$ (1 orbit, right)

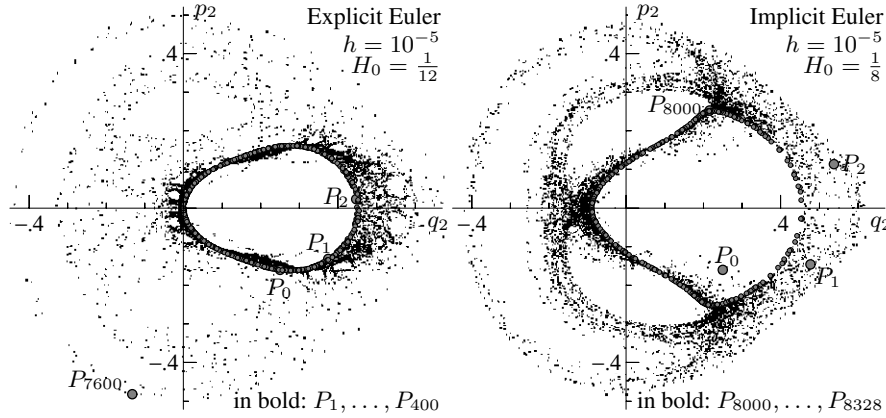


Fig. 3.3. Poincaré cuts for numerical methods, one orbit each; explicit Euler (left), implicit Euler (right). Same initial data as in Fig. 3.2

$$\ddot{q}_1 = -q_1 - 2q_1q_2, \quad \ddot{q}_2 = -q_2 - q_1^2 + q_2^2 \quad (3.3)$$

has solutions with nontrivial properties. For given initial values with $H(p_0, q_0) < \frac{1}{6}$ and q_0 inside the triangle $U \leq \frac{1}{6}$, the solution stays there and moves somehow like a mass point gliding on this surface (see Fig. 3.1, right).

Poincaré Cuts. We fix first the energy H_0 and put $q_{10} = 0$. Then for any point $P_0 = (q_{20}, p_{20})$, we obtain p_{10} from (3.1) as $p_{10} = \sqrt{2H_0 - 2U_0 - p_{20}^2}$, where we choose the positive root. We then follow the solution until it hits again the surface $q_1 = 0$ in the positive direction $p_1 > 0$ and obtain a point $P_1 = (q_{21}, p_{21})$; in the same way we compute $P_2 = (q_{22}, p_{22})$, etc. For the same initial values as in Fig. 3.1 and with $H_0 = \frac{1}{12}$, the solution for $0 \leq t \leq 300\,000$ gives 46 865 Poincaré cuts which are all displayed in Fig. 3.2 (left). They seem to lie exactly on a curve, as do the orbits for 5 other choices of initial values. This picture thus shows “convincing

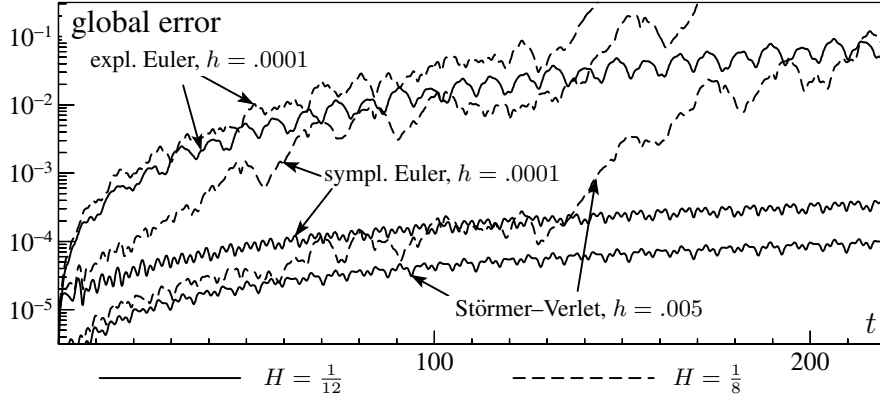


Fig. 3.4. Global error of numerical methods for nearly quasiperiodic and for chaotic solutions; same initial data as in Fig. 3.2

evidence” for the existence of a second invariant, for which Gustavson (1966) has derived a formal expansion, whose first terms represent perfectly these curves.

“But here comes the surprise” (Hénon–Heiles, p. 76): Fig. 3.2 shows to the right the same picture in the (q_2, p_2) plane for a somewhat higher Energy $H = \frac{1}{8}$. The motion turns completely to chaos and all hope for a second invariant disappears. Actually, Gustavson’s series does not converge.

Numerical Experiments. We now apply numerical methods, the *explicit* Euler method to the low energy initial values $H = \frac{1}{12}$ (Fig. 3.3, left), and the *implicit* Euler method to the high energy initial values (Fig. 3.3, right), both methods with a very small step size $h = 10^{-5}$. As we already expect from our previous experiences, the explicit Euler method tends to *increase* the energy and turns order into chaos, while the implicit Euler method tends to *decrease* it and turns chaos into order. The Störmer–Verlet method (not shown) behaves as the exact solution even for step sizes as large as $h = 10^{-1}$.

In our next experiment we study the *global error* (see Fig. 3.4), once for the case of the nearly quasiperiodic orbit ($H = \frac{1}{12}$) and once for the chaotic one ($H = \frac{1}{8}$), both for the explicit Euler, the symplectic Euler, and the Störmer–Verlet scheme. It may come as a surprise, that only in the first case we have the same behaviour (linear or quadratic growth) as in Fig. 2.3 for the Kepler problem. In the second case ($H = \frac{1}{8}$) the global error grows exponentially for all methods, and the explicit Euler method is worst.

Study of a Mapping. The passage from a point P_i to the next one P_{i+1} (as explained for the left picture of Fig. 3.2) can be considered as a *mapping* $\Phi : P_i \mapsto P_{i+1}$ and the sequence of points P_0, P_1, P_2, \dots are just the iterates of this mapping. This mapping is represented for the two energy levels $H = \frac{1}{12}$ and $H = \frac{1}{8}$ in Fig. 3.5 and its study allows to better understand the behaviour of the orbits. We see no significant difference between the two cases, simply for larger H the deformations are more violent and correspond to larger eigenvalues of the Jacobian of Φ . In

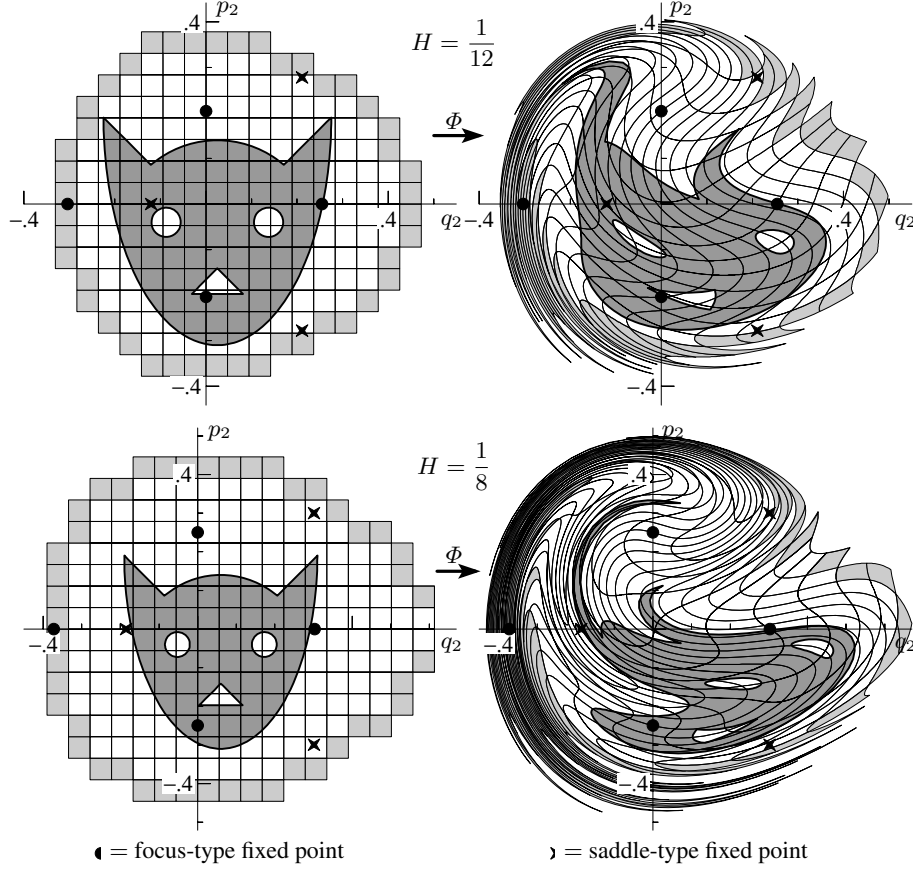


Fig. 3.5. The Poincaré map $\Phi : P_0 \rightarrow P_1$ for the Hénon-Heiles Model

both cases we have seven fixed points, which correspond to periodic solutions of the system (3.3). Four of them are stable and lie inside the white islands of Fig. 3.2.

I.4 Molecular Dynamics

We do not need exact classical trajectories to do this, but must lay great emphasis on energy conservation as being of primary importance for this reason.
(M.P. Allen & D.J. Tildesley 1987)

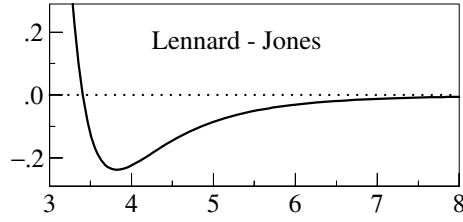
Molecular dynamics requires the solution of Hamiltonian systems (1.10), where the total energy is given by

$$H(p, q) = \frac{1}{2} \sum_{i=1}^N \frac{1}{m_i} p_i^T p_i + \sum_{i=2}^N \sum_{j=1}^{i-1} V_{ij}(\|q_i - q_j\|), \quad (4.1)$$

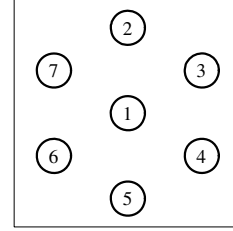
and $V_{ij}(r)$ are given potential functions. Here, q_i and p_i denote the positions and momenta of atoms and m_i is the atomic mass of the i th atom. We remark that the outer solar system (2.12) is such an N -body system with $V_{ij}(r) = -Gm_i m_j / r$. In molecular dynamics the Lennard–Jones potential

$$V_{ij}(r) = 4\varepsilon_{ij} \left(\left(\frac{\sigma_{ij}}{r} \right)^{12} - \left(\frac{\sigma_{ij}}{r} \right)^6 \right) \quad (4.2)$$

is very popular (ε_{ij} and σ_{ij} are suitable constants depending on the atoms). This potential has an absolute minimum at distance $r = \sigma_{ij} \sqrt[6]{2}$. The force due to this potential strongly repels the atoms when they are closer than this value, and they attract each other when they are farther away.



Numerical Experiments with a Frozen Argon Crystal. As in Biesiadecki & Skeel (1993) we consider the interaction of seven argon atoms in a plane, where six of them are arranged symmetrically around a centre atom. As a mathematical model we take the Hamiltonian (4.1) with $N = 7$, $m_i = m = 66.34 \cdot 10^{-27}$ [kg],



$$\varepsilon_{ij} = \varepsilon = 119.8 k_B [\text{J}], \quad \sigma_{ij} = \sigma = 0.341 [\text{nm}],$$

where $k_B = 1.380658 \cdot 10^{-23}$ [J/K] is Boltzmann's constant (see Allen & Tildesley (1987), page 21). As units for our calculations we take masses in [kg], distances in nanometers ($1 [\text{nm}] = 10^{-9} [\text{m}]$), and times in nanoseconds ($1 [\text{nsec}] = 10^{-9} [\text{sec}]$). Initial positions (in [nm]) and initial velocities (in [nm/nsec]) are given in Table 4.1. They are chosen such that neighbouring atoms have a distance that is close to the one with lowest potential energy, and such that the total momentum is zero and therefore the centre of gravity does not move. The energy at the initial position is $H(p_0, q_0) \approx -1260.2 k_B [\text{J}]$.

For computations in molecular dynamics one is usually not interested in the trajectories of the atoms, but one aims at macroscopic quantities such as temperature, pressure, internal energy, etc. Here we consider the total energy, given by the Hamiltonian, and the temperature which can be calculated from the formula (see Allen &

Table 4.1. Initial values for the simulation of a frozen argon crystal

atom	1	2	3	4	5	6	7
position	0.00	0.02	0.34	0.36	-0.02	-0.35	-0.31
	0.00	0.39	0.17	-0.21	-0.40	-0.16	0.21
velocity	-30	50	-70	90	80	-40	-80
	-20	-90	-60	40	90	100	-60

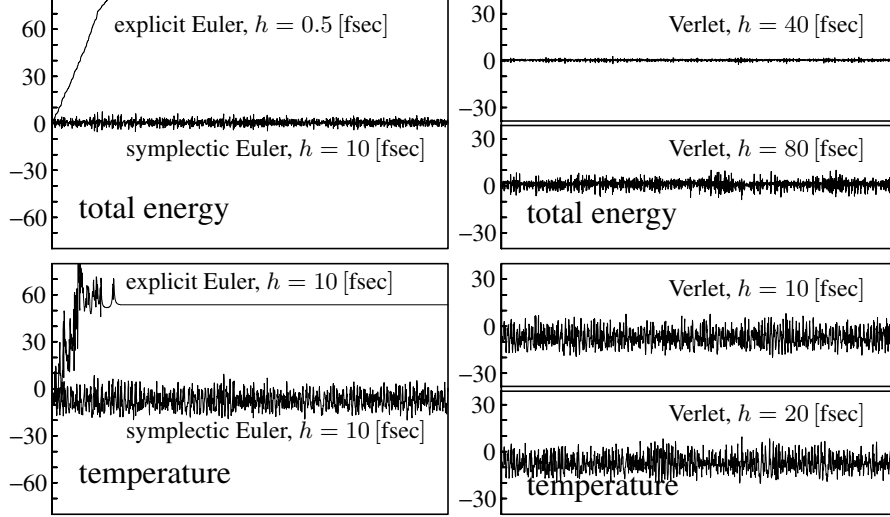


Fig. 4.1. Computed total energy and temperature of the argon crystal

Tildesley (1987), page 46)

$$T = \frac{1}{2Nk_B} \sum_{i=1}^N m_i \|\dot{q}_i\|^2. \quad (4.3)$$

We apply the explicit and symplectic Euler methods and also the Verlet method to this problem. Observe that for a Hamiltonian such as (4.1) all three methods are explicit, and all of them need only one force evaluation per integration step. In Fig. 4.1 we present the numerical results of our experiments. The integrations are done over an interval of length 0.2 [nsec]. The step sizes are indicated in femtoseconds (1 [fsec] = 10^{-6} [nsec]).

The two upper pictures show the values $(H(p_n, q_n) - H(p_0, q_0))/k_B$ as a function of time $t_n = nh$. For the exact solution, this value is precisely zero for all times. Similar to earlier experiments we see that the symplectic Euler method is qualitatively correct, whereas the numerical solution of the explicit Euler method, although computed with a much smaller step size, is completely useless (see the citation at the beginning of this section). The Verlet method is qualitatively correct and gives much more accurate results than the symplectic Euler method (we shall see later that the Verlet method is of order 2). The two computations with the Verlet method show that the energy error decreases by a factor of 4 if the step size is reduced by a factor of 2 (second order convergence).

The two lower pictures of Fig. 4.1 show the numerical values of the temperature difference $T - T_0$ with T given by (4.3) and $T_0 \approx 22.72$ [K] (initial temperature). In contrast to the total energy, this is not an exact invariant, but for our problem it fluctuates around a constant value. The explicit Euler method gives wrong results,

but the symplectic Euler and the Verlet methods show the desired behaviour. This time a reduction of the step size does not reduce the amplitude of the oscillations, which indicates that the fluctuation of the exact temperature is of the same size.

I.5 Highly Oscillatory Problems

In this section we discuss a system with almost-harmonic high-frequency oscillations. We show numerical phenomena of methods applied with step sizes that are not small compared to the period of the fastest oscillations.

I.5.1 A Fermi–Pasta–Ulam Problem

... dealing with the behavior of certain nonlinear physical systems where the non-linearity is introduced as a perturbation to a primarily linear problem. The behavior of the systems is to be studied for times which are long compared to the characteristic periods of the corresponding linear problems. (E. Fermi, J. Pasta, S. Ulam 1955)

In the early 1950s MANIAC-I had just been completed and sat poised for an attack on significant problems. ... Fermi suggested that it would be highly instructive to integrate the equations of motion numerically for a judiciously chosen, one-dimensional, harmonic chain of mass points weakly perturbed by nonlinear forces. (J. Ford 1992)

The problem of Fermi, Pasta & Ulam (1955) is a simple model for simulations in statistical mechanics which revealed highly unexpected dynamical behaviour. We consider a modification consisting of a chain of $2m$ mass points, connected with alternating soft nonlinear and stiff linear springs, and fixed at the end points (see Gagliani, Giorgilli, Martinoli & Vanzini (1992) and Fig. 5.1). The variables q_1, \dots, q_{2m}

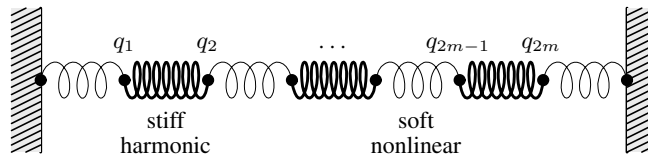


Fig. 5.1. Chain with alternating soft nonlinear and stiff linear springs

($q_0 = q_{2m+1} = 0$) stand for the displacements of the mass points, and $p_i = \dot{q}_i$ for their velocities. The motion is described by a Hamiltonian system with total energy

$$H(p, q) = \frac{1}{2} \sum_{i=1}^m (p_{2i-1}^2 + p_{2i}^2) + \frac{\omega^2}{4} \sum_{i=1}^m (q_{2i} - q_{2i-1})^2 + \sum_{i=0}^m (q_{2i+1} - q_{2i})^4,$$

where ω is assumed to be large. It is quite natural to introduce the new variables

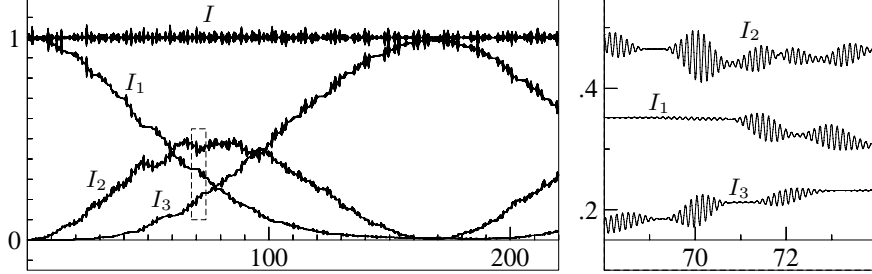


Fig. 5.2. Exchange of energy in the exact solution of the Fermi-Pasta-Ulam model. The picture to the right is an enlargement of the narrow rectangle in the left-hand picture

$$\begin{aligned} x_{0,i} &= (q_{2i} + q_{2i-1})/\sqrt{2}, & x_{1,i} &= (q_{2i} - q_{2i-1})/\sqrt{2}, \\ y_{0,i} &= (p_{2i} + p_{2i-1})/\sqrt{2}, & y_{1,i} &= (p_{2i} - p_{2i-1})/\sqrt{2}, \end{aligned} \quad (5.1)$$

where $x_{0,i}$ ($i = 1, \dots, m$) represents a scaled displacement of the i th stiff spring, $x_{1,i}$ a scaled expansion (or compression) of the i th stiff spring, and $y_{0,i}, y_{1,i}$ their velocities (or momenta). With this change of coordinates, the motion in the new variables is again described by a Hamiltonian system, with

$$\begin{aligned} H(y, x) &= \frac{1}{2} \sum_{i=1}^m (y_{0,i}^2 + y_{1,i}^2) + \frac{\omega^2}{2} \sum_{i=1}^m x_{1,i}^2 + \frac{1}{4} \left((x_{0,1} - x_{1,1})^4 + \right. \\ &\quad \left. + \sum_{i=1}^{m-1} (x_{0,i+1} - x_{1,i+1} - x_{0,i} - x_{1,i})^4 + (x_{0,m} + x_{1,m})^4 \right). \end{aligned} \quad (5.2)$$

Besides the fact that the equations of motion are Hamiltonian, so that the total energy is exactly conserved, they have a further interesting feature. Let

$$I_j(x_{1,j}, y_{1,j}) = \frac{1}{2} (y_{1,j}^2 + \omega^2 x_{1,j}^2) \quad (5.3)$$

denote the energy of the j th stiff spring. It turns out that there is an exchange of energy between the stiff springs, but the total oscillatory energy $I = I_1 + \dots + I_m$ remains close to a constant value, in fact, $I((x(t), y(t))) = I((x(0), y(0))) + \mathcal{O}(\omega^{-1})$. For an illustration of this property, we choose $m = 3$ (as in Fig. 5.1), $\omega = 50$,

$$x_{0,1}(0) = 1, \quad y_{0,1}(0) = 1, \quad x_{1,1}(0) = \omega^{-1}, \quad y_{1,1}(0) = 1,$$

and zero for the remaining initial values. Fig. 5.2 displays the energies I_1, I_2, I_3 of the stiff springs together with the total oscillatory energy $I = I_1 + I_2 + I_3$ as a function of time. The solution has been computed very carefully with high accuracy, so that the displayed oscillations can be considered as exact.

I.5.2 Application of Classical Integrators

Which of the methods of the foregoing sections produce qualitatively correct approximations when the product of the step size h with the high frequency ω is relatively large?

Linear Stability Analysis. To get an idea of the maximum admissible step size, we neglect the quartic term in the Hamiltonian (5.2), so that the differential equation splits into the two-dimensional problems $\dot{y}_{0,i} = 0$, $\dot{x}_{0,i} = y_{0,i}$ and

$$\dot{y}_{1,i} = -\omega^2 x_{1,i}, \quad \dot{x}_{1,i} = y_{1,i}. \quad (5.4)$$

Omitting the subscripts, the solution of (5.4) is

$$\begin{pmatrix} y(t) \\ \omega x(t) \end{pmatrix} = \begin{pmatrix} \cos \omega t & -\sin \omega t \\ \sin \omega t & \cos \omega t \end{pmatrix} \begin{pmatrix} y(0) \\ \omega x(0) \end{pmatrix}.$$

The numerical solution of a one-step method applied to (5.4) yields

$$\begin{pmatrix} y_{n+1} \\ \omega x_{n+1} \end{pmatrix} = M(h\omega) \begin{pmatrix} y_n \\ \omega x_n \end{pmatrix}, \quad (5.5)$$

and the eigenvalues λ_i of $M(h\omega)$ determine the long-time behaviour of the numerical solution. Stability (i.e., boundedness of the solution of (5.5)) requires the eigenvalues to be less than or equal to one in modulus. For the explicit Euler method we have $\lambda_{1,2} = 1 \pm ih\omega$, so that the energy $I_n = (y_n^2 + \omega^2 x_n^2)/2$ increases as $(1 + h^2\omega^2)^{n/2}$. For the implicit Euler method we have $\lambda_{1,2} = (1 \pm ih\omega)^{-1}$, and the energy decreases as $(1 + h^2\omega^2)^{-n/2}$. For the implicit midpoint rule, the matrix $M(h\omega)$ is orthogonal and therefore I_n is exactly preserved for all h and for all times. Finally, for the symplectic Euler method and for the Störmer–Verlet scheme we have

$$M(h\omega) = \begin{pmatrix} 1 & -h\omega \\ h\omega & 1 - h^2\omega^2 \end{pmatrix}, \quad M(h\omega) = \begin{pmatrix} 1 - \frac{h^2\omega^2}{2} & -\frac{h\omega}{2} \left(1 - \frac{h^2\omega^2}{4}\right) \\ \frac{h\omega}{2} & 1 - \frac{h^2\omega^2}{2} \end{pmatrix},$$

respectively. For both matrices, the characteristic polynomial is $\lambda^2 - (2 - h^2\omega^2)\lambda + 1$, so that the eigenvalues are of modulus one if and only if $|h\omega| \leq 2$.

Numerical Experiments. We apply several methods to the Fermi–Pasta–Ulam (FPU) problem, with $\omega = 50$ and initial data as given in Sect. I.5.1. The explicit and implicit Euler methods give completely wrong solutions even for very small step sizes. Fig. 5.3 presents the numerical results for H , I , I_1 , I_2 , I_3 obtained with the implicit midpoint rule, the symplectic Euler, and the Störmer–Verlet scheme. For the small step size $h = 0.001$ all methods give satisfactory results, although the energy exchange is not reproduced accurately over long times. The Hamiltonian H and the total oscillatory energy I are well conserved over much longer time intervals. The larger step size $h = 0.03$ has been chosen such that $h\omega = 1.5$ is close

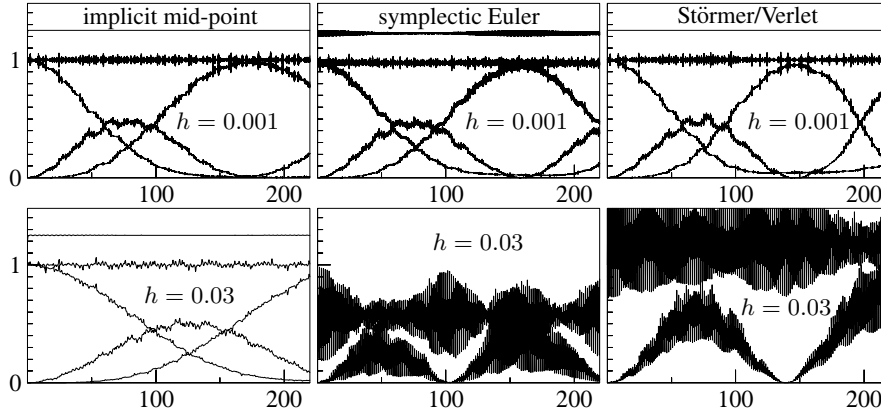


Fig. 5.3. Numerical solution for the FPU problem (5.2) with data as in Sect. I.5.1, obtained with the implicit midpoint rule (left), symplectic Euler (middle), and Störmer–Verlet scheme (right); the upper pictures use $h = 0.001$, the lower pictures $h = 0.03$; the first four pictures show the Hamiltonian $H - 0.8$ and the oscillatory energies I_1, I_2, I_3, I ; the last two pictures only show I_2 and I

to the stability limit of the symplectic Euler and the Störmer–Verlet methods. The values of H and I are still bounded over very long time intervals, but the oscillations do not represent the true behaviour. Moreover, the average value of I is no longer close to 1, as it is for the exact solution. These phenomena call for an explanation, and for numerical methods with an improved behaviour (see Chap. XIII).

I.6 Exercises

1. Show that the Lotka–Volterra problem (1.1) in logarithmic scale, i.e., by putting $p = \log u$ and $q = \log v$, becomes a Hamiltonian system with the function (1.4) as Hamiltonian (see Fig. 6.1).

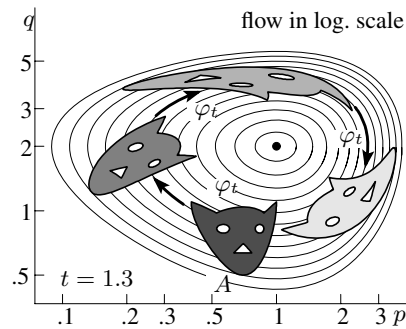


Fig. 6.1. Area preservation in logarithmic scale of the Lotka–Volterra flow

2. Apply the symplectic Euler method (or the implicit midpoint rule) to problems such as

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} (v-2)/v \\ (1-u)/u \end{pmatrix}, \quad \begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} u^2 v(v-2) \\ v^2 u(1-u) \end{pmatrix}$$

with various initial conditions. Both problems have the same first integral (1.4) as the Lotka–Volterra problem and therefore their solutions are also periodic. Do the numerical solutions also show this behaviour?

3. A general two-body problem (sun and planet) is given by the Hamiltonian

$$H(p, p_S, q, q_S) = \frac{1}{2M} p_S^T p_S + \frac{1}{2m} p^T p - \frac{GmM}{\|q - q_S\|},$$

where $q_S, q \in \mathbb{R}^3$ are the positions of the sun (mass M) and the planet (mass m), $p_S, p \in \mathbb{R}^3$ are their momenta, and G is the gravitational constant.

- a) Prove: in heliocentric coordinates $Q := q - q_S$, the equations of motion are

$$\ddot{Q} = -G(M+m) \frac{Q}{\|Q\|^3}.$$

- b) Prove that $\frac{d}{dt}(Q(t) \times \dot{Q}(t)) = 0$, so that $Q(t)$ stays for all times t in the plane $E = \{q; d^T q = 0\}$, where $d = Q(0) \times \dot{Q}(0)$.

Conclusion. The coordinates corresponding to a basis in E satisfy the two-dimensional equations (2.2).

4. In polar coordinates, the two-body problem (2.2) becomes

$$\ddot{r} = -V'(r) \quad \text{with} \quad V(r) = \frac{L_0^2}{2r^2} - \frac{1}{r}$$

which is independent of φ . The angle $\varphi(t)$ can be obtained by simple integration from $\dot{\varphi}(t) = L_0/r^2(t)$.

5. Compute the period of the solution of the Kepler problem (2.2) and deduce from the result Kepler's "third law".

Hint. Comparing Kepler's second law (2.6) with the area of the ellipse gives $\frac{1}{2} L_0 T = ab\pi$. Then apply (2.7). The result is $T = 2\pi(2|H_0|)^{-3/2} = 2\pi a^3/2$.

6. Deduce Kepler's first law from (2.2) by the elegant method of Laplace (1799).

Hint. Multiplying (2.2) with (2.5) gives

$$L_0 \ddot{q}_1 = \frac{d}{dt} \left(\frac{q_2}{r} \right), \quad L_0 \ddot{q}_2 = \frac{d}{dt} \left(-\frac{q_1}{r} \right),$$

and after integration $L_0 \dot{q}_1 = \frac{q_2}{r} + B$, $L_0 \dot{q}_2 = -\frac{q_1}{r} + A$, where A and B are integration constants. Then eliminate \dot{q}_1 and \dot{q}_2 by multiplying these equations by q_2 and $-q_1$ respectively and by subtracting them. The result is a quadratic equation in q_1 and q_2 .

7. Whatever the initial values for the Kepler problem are, $1 + 2H_0 L_0^2 \geq 0$ holds. Hence, the value e is well defined by (2.9).

Hint. L_0 is the area of the parallelogram spanned by the vectors $q(0)$ and $\dot{q}(0)$.

8. *Implementation of the Störmer–Verlet scheme.* Explain why the use of the one-step formulation (1.17) is numerically more stable than that of the two-term recursion (1.15).
9. *Runge–Lenz–Pauli vector.* Prove that the function

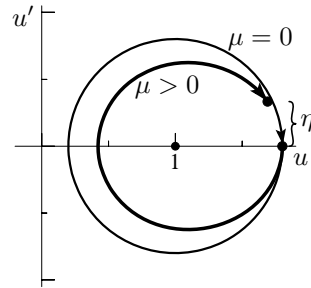
$$A(p, q) = \begin{pmatrix} p_1 \\ p_2 \\ 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ 0 \\ q_1 p_2 - q_2 p_1 \end{pmatrix} - \frac{1}{\sqrt{q_1^2 + q_2^2}} \begin{pmatrix} q_1 \\ q_2 \\ 0 \end{pmatrix}$$

is a first integral of the Kepler problem, i.e., $A(p(t), q(t)) = \text{Const}$ along solutions of the problem. However, it is not a first integral of the perturbed Kepler problem of Exercise 12.

10. Add a column to Table 2.1 which shows the long-time behaviour of the error in the Runge–Lenz–Pauli vector (see Exercise 9) for the various numerical integrators.
11. For the Kepler problem, eliminate (p_1, p_2) from the relations $H(p, q) = \text{Const}$, $L(p, q) = \text{Const}$ and $A(p, q) = \text{Const}$. This gives a quadratic relation for (q_1, q_2) and proves that the solution lies on an ellipse, a parabola, or on a hyperbola.
12. Study numerically the solution of the perturbed Kepler problem with Hamiltonian

$$H(p_1, p_2, q_1, q_2) = \frac{1}{2} (p_1^2 + p_2^2) - \frac{1}{\sqrt{q_1^2 + q_2^2}} - \frac{\mu}{3\sqrt{(q_1^2 + q_2^2)^3}},$$

where μ is a positive or negative small number. Among others, this problem describes the motion of a planet in the Schwarzschild potential for Einstein's general relativity theory⁷. You will observe a precession of the perihelion, which, applied to the orbit of Mercury, represented the historically first verification of Einstein's theory (see e.g., Birkhoff 1923, p. 261-264).



The precession can also be expressed analytically: the equation for $u = 1/r$ as a function of φ , corresponding to (2.8), here becomes

$$u'' + u = \frac{1}{d} + \mu u^2, \quad (6.1)$$

where $d = L_0^2$. Now compute the derivative of this solution with respect to μ , at $\mu = 0$ and $u = (1 + e \cos(\varphi - \varphi^*)) / d$ after one period $t = 2\pi$. This leads to $\eta = \mu(e/d^2) \cdot 2\pi \sin \varphi$ (see the small picture). Then, for small μ , the precession after one period is

$$\Delta\varphi = \frac{2\pi\mu}{d}. \quad (6.2)$$

⁷ We are grateful to Prof. Ruth Durrer for helpful hints about this subject.

Chapter II.

Numerical Integrators

After having seen in Chap. I some simple numerical methods and a variety of numerical phenomena that they exhibited, we now present more elaborate classes of numerical methods. We start with Runge–Kutta and collocation methods, and we introduce discontinuous collocation methods, which cover essentially all high-order implicit Runge–Kutta methods of interest. We then treat partitioned Runge–Kutta methods and Nyström methods, which can be applied to partitioned problems such as Hamiltonian systems. Finally we present composition and splitting methods.

II.1 Runge–Kutta and Collocation Methods

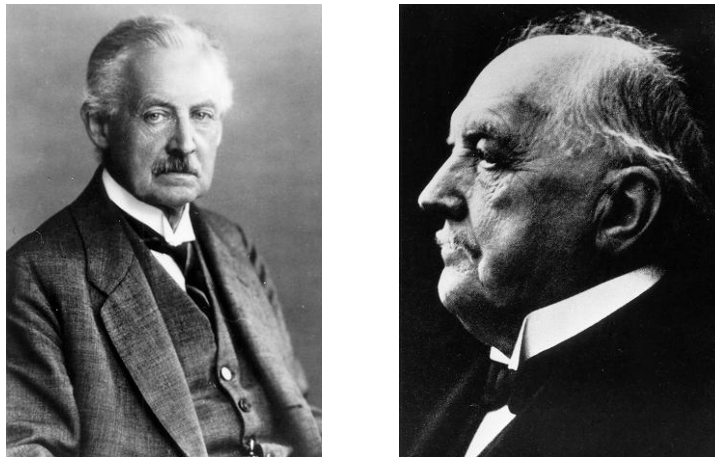


Fig. 1.1. Carl David Tolmé Runge (left picture), born: 30 August 1856 in Bremen (Germany), died: 3 January 1927 in Göttingen (Germany). Wilhelm Martin Kutta (right picture), born: 3 November 1867 in Pitschen, Upper Silesia (now Byczyna, Poland), died: 25 December 1944 in Fürstenfeldbruck (Germany)

Runge–Kutta methods form an important class of methods for the integration of differential equations. A special subclass, the collocation methods, allows for a particularly elegant access to order, symplecticity and continuous output.

II.1.1 Runge–Kutta Methods

In this section, we treat non-autonomous systems of first-order ordinary differential equations

$$\dot{y} = f(t, y), \quad y(t_0) = y_0. \quad (1.1)$$

The integration of this equation gives $y(t_1) = y_0 + \int_{t_0}^{t_1} f(t, y(t)) dt$, and replacing the integral by the trapezoidal rule, we obtain

$$y_1 = y_0 + \frac{h}{2} (f(t_0, y_0) + f(t_1, y_1)). \quad (1.2)$$

This is the *implicit trapezoidal rule*, which, in addition to its historical importance for computations in partial differential equations (Crank–Nicolson) and in A-stability theory (Dahlquist), played a crucial role even earlier in the discovery of Runge–Kutta methods. It was the starting point of Runge (1895), who “predicted” the unknown y_1 -value to the right by an Euler step, and obtained the first of the following formulas (the second being the analogous formula for the midpoint rule)

$$\begin{aligned} k_1 &= f(t_0, y_0) & k_1 &= f(t_0, y_0) \\ k_2 &= f(t_0 + h, y_0 + hk_1) & k_2 &= f(t_0 + \frac{h}{2}, y_0 + \frac{h}{2}k_1) \\ y_1 &= y_0 + \frac{h}{2}(k_1 + k_2) & y_1 &= y_0 + hk_2. \end{aligned} \quad (1.3)$$

These methods have a nice geometric interpretation (which is illustrated in the first two pictures of Fig. 1.2 for a famous problem, the Riccati equation): they consist of polygonal lines, which assume the slopes prescribed by the differential equation evaluated at previous points.

Idea of Heun (1900) and Kutta (1901): compute several polygonal lines, each starting at y_0 and assuming the various slopes k_j on portions of the integration interval, which are proportional to some given constants a_{ij} ; at the final point of each polygon evaluate a new slope k_i . The last of these polygons, with constants b_i , determines the numerical solution y_1 (see the third picture of Fig. 1.2). This idea leads to the class of *explicit* Runge–Kutta methods, i.e., formula (1.4) below with $a_{ij} = 0$ for $i \leq j$.

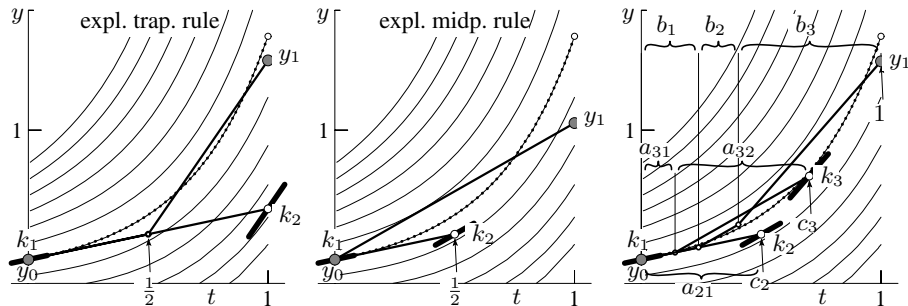


Fig. 1.2. Runge–Kutta methods for $\dot{y} = t^2 + y^2$, $y_0 = 0.46$, $h = 1$; dotted: exact solution

Much more important for our purpose are *implicit* Runge–Kutta methods, introduced mainly in the work of Butcher (1963).

Definition 1.1. Let b_i, a_{ij} ($i, j = 1, \dots, s$) be real numbers and let $c_i = \sum_{j=1}^s a_{ij}$. An s -stage Runge–Kutta method is given by

$$\begin{aligned} k_i &= f\left(t_0 + c_i h, y_0 + h \sum_{j=1}^s a_{ij} k_j\right), \quad i = 1, \dots, s \\ y_1 &= y_0 + h \sum_{i=1}^s b_i k_i. \end{aligned} \quad (1.4)$$

Here we allow a full matrix (a_{ij}) of non-zero coefficients. In this case, the slopes k_i can no longer be computed explicitly, and even do not necessarily exist. For example, for the problem set-up of Fig. 1.2 the implicit trapezoidal rule has no solution. However, the implicit function theorem assures that, for sufficiently small h , the nonlinear system (1.4) for the values k_1, \dots, k_s has a locally unique solution close to $k_i \approx f(t_0, y_0)$.

Since Butcher's work, the coefficients are usually displayed as follows:

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array}. \quad (1.5)$$

Definition 1.2. A Runge–Kutta method (or a general one-step method) has *order* p , if for all sufficiently regular problems (1.1) the *local error* $y_1 - y(t_0 + h)$ satisfies

$$y_1 - y(t_0 + h) = \mathcal{O}(h^{p+1}) \quad \text{as } h \rightarrow 0.$$

To check the order of a Runge Kutta method, one has to compute the Taylor series expansions of $y(t_0 + h)$ and y_1 around to $h = 0$. This leads to the following algebraic conditions for the coefficients for orders 1, 2, and 3:

$$\begin{aligned} & \sum_i b_i = 1 && \text{for order 1;} \\ \text{in addition} & \sum_i b_i c_i = 1/2 && \text{for order 2;} \\ \text{in addition} & \sum_i b_i c_i^2 = 1/3 && \\ \text{and} & \sum_{i,j} b_i a_{ij} c_j = 1/6 && \text{for order 3.} \end{aligned} \quad (1.6)$$

For higher orders, however, this problem represented a great challenge in the first half of the 20th century. We shall present an elegant theory in Sect. III.1 which allows order conditions to be derived.

Among the methods seen up to now, the explicit and implicit Euler methods

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array} \quad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \quad (1.7)$$

are of order 1, the implicit trapezoidal and midpoint rules as well as both methods of Runge

$$\begin{array}{c|cc} 0 & & \\ \hline 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \quad \begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array} \quad \begin{array}{c|cc} 0 & & \\ \hline 1 & 1 & \\ \hline & 1/2 & 1/2 \end{array} \quad \begin{array}{c|cc} 0 & & \\ \hline 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array}$$

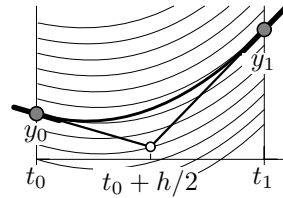
are of order 2. The most successful methods during more than half a century were the 4th order methods of Kutta:

$$\begin{array}{c|cccc} 0 & & & & \\ \hline 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 2/6 & 2/6 & 1/6 \end{array} \quad \begin{array}{c|cccc} 0 & & & & \\ \hline 1/3 & 1/3 & & & \\ 2/3 & -1/3 & 1 & & \\ 1 & 1 & -1 & 1 & \\ \hline & 1/8 & 3/8 & 3/8 & 1/8 \end{array} \quad (1.8)$$

II.1.2 Collocation Methods

The high speed computing machines make it possible to enjoy the advantages of intricate methods. (P.C. Hammer & J.W. Hollingsworth 1955)

Collocation methods for ordinary differential equations have their origin, once again, in the implicit trapezoidal rule (1.2): Hammer & Hollingsworth (1955) discovered that this method can be interpreted as being generated by a *quadratic function* “which agrees in direction with that indicated by the differential equation at two points” t_0 and t_1 (see the picture to the right). This idea allows one to “see much-used methods in a new light” and allows various generalizations (Guillou & Soulé (1969), Wright (1970)). An interesting feature of collocation methods is that we not only get a discrete set of approximations, but also a *continuous approximation* to the solution.



Definition 1.3. Let c_1, \dots, c_s be distinct real numbers (usually $0 \leq c_i \leq 1$). The *collocation polynomial* $u(t)$ is a polynomial of degree s satisfying

$$\begin{aligned} u(t_0) &= y_0 \\ \dot{u}(t_0 + c_i h) &= f(t_0 + c_i h, u(t_0 + c_i h)), \quad i = 1, \dots, s, \end{aligned} \quad (1.9)$$

and the numerical solution of the *collocation method* is defined by $y_1 = u(t_0 + h)$.

For $s = 1$, the polynomial has to be of the form $u(t) = y_0 + (t - t_0)k$ with

$$k = f(t_0 + c_1 h, y_0 + h c_1 k).$$

We see that the explicit and implicit Euler methods and the midpoint rule are collocation methods with $c_1 = 0$, $c_1 = 1$ and $c_1 = 1/2$, respectively.

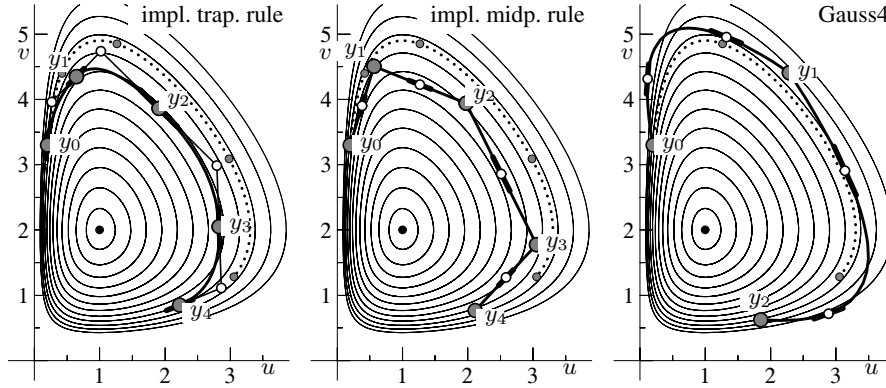
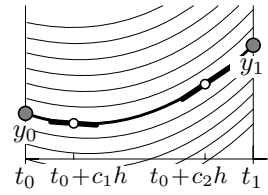


Fig. 1.3. Collocation solutions for the Lotka–Volterra problem (I.1.1); $u_0 = 0.2$, $v_0 = 3.3$; methods of order 2: four steps with $h = 0.4$; method of order 4: two steps with $h = 0.8$; dotted: exact solution

For $s = 2$ and $c_1 = 0, c_2 = 1$ we find, of course, the implicit trapezoidal rule. The choice of Hammer & Hollingsworth for the collocation points is $c_{1,2} = 1/2 \pm \sqrt{3}/6$, the *Gaussian quadrature nodes* (see the picture to the right). We will see that the corresponding method is of order 4.



In Fig. 1.3 we illustrate the collocation idea with these methods for the Lotka–Volterra problem (I.1.1). One can observe that, in spite of the extremely large step sizes, the methods are quite satisfactory.

Theorem 1.4 (Guillou & Soulé 1969, Wright 1970). *The collocation method of Definition 1.3 is equivalent to the s -stage Runge–Kutta method (1.4) with coefficients*

$$a_{ij} = \int_0^{c_i} \ell_j(\tau) d\tau, \quad b_i = \int_0^1 \ell_i(\tau) d\tau, \quad (1.10)$$

where $\ell_i(\tau)$ is the Lagrange polynomial $\ell_i(\tau) = \prod_{l \neq i} (\tau - c_l) / (c_i - c_l)$.

Proof. Let $u(t)$ be the collocation polynomial and define

$$k_i := \dot{u}(t_0 + c_i h).$$

By the Lagrange interpolation formula we have $\dot{u}(t_0 + \tau h) = \sum_{j=1}^s k_j \cdot \ell_j(\tau)$, and by integration we get

$$u(t_0 + c_i h) = y_0 + h \sum_{j=1}^s k_j \int_0^{c_i} \ell_j(\tau) d\tau.$$

Inserted into (1.9) this gives the first formula of the Runge–Kutta equation (1.4). Integration from 0 to 1 yields the second one. \square

The above proof can also be read in reverse order. This shows that a Runge–Kutta method with coefficients given by (1.10) can be interpreted as a collocation method. Since $\tau^{k-1} = \sum_{j=1}^s c_j^{k-1} \ell_j(\tau)$ for $k = 1, \dots, s$, the relations (1.10) are equivalent to the linear systems

$$\begin{aligned} C(q) : \quad & \sum_{j=1}^s a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \quad k = 1, \dots, q, \quad \text{all } i \\ B(p) : \quad & \sum_{i=1}^s b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, \dots, p, \end{aligned} \quad (1.11)$$

with $q = s$ and $p = s$. What is the order of a Runge–Kutta method whose coefficients b_i, a_{ij} are determined in this way?

Compared to the enormous difficulties that the first explorers had in constructing Runge–Kutta methods of orders 5 and 6, and also compared to the difficult algebraic proofs of the first papers of Butcher, the following general theorem and its proof, discovered in this form by Guillou & Soulé (1969), are surprisingly simple.

Theorem 1.5 (Superconvergence). *If the condition $B(p)$ holds for some $p \geq s$, then the collocation method (Definition 1.3) has order p . This means that the collocation method has the same order as the underlying quadrature formula.*

Proof. We consider the collocation polynomial $u(t)$ as the solution of a perturbed differential equation

$$\dot{u} = f(t, u) + \delta(t) \quad (1.12)$$

with defect $\delta(t) := \dot{u}(t) - f(t, u(t))$. Subtracting (1.1) from (1.12) we get after linearization that

$$\dot{u}(t) - \dot{y}(t) = \frac{\partial f}{\partial y}(t, y(t)) (u(t) - y(t)) + \delta(t) + r(t), \quad (1.13)$$

where, for $t_0 \leq t \leq t_0 + h$, the remainder $r(t)$ is of size $\mathcal{O}(\|u(t) - y(t)\|^2) = \mathcal{O}(h^{2s+2})$ by Lemma 1.6 below. The variation of constants formula (see e.g., Hairer, Nørsett & Wanner (1993), p. 66) then yields

$$y_1 - y(t_0 + h) = u(t_0 + h) - y(t_0 + h) = \int_{t_0}^{t_0+h} R(t_0 + h, s) (\delta(s) + r(s)) ds, \quad (1.14)$$

where $R(t, s)$ is the resolvent of the homogeneous part of the differential equation (1.13), i.e., the solution of the matrix differential equation $\partial R(t, s)/\partial t = A(t)R(t, s)$, $R(s, s) = I$, with $A(t) = \partial f/\partial y(t, y(t))$. The integral over $R(t_0 + h, s)r(s)$ gives a $\mathcal{O}(h^{2s+3})$ contribution. The main idea now is to apply the quadrature formula $(b_i, c_i)_{i=1}^s$ to the integral over $g(s) = R(t_0 + h, s)\delta(s)$; because the defect $\delta(s)$ vanishes at the collocation points $t_0 + c_i h$ for $i = 1, \dots, s$, this gives zero as the numerical result. Thus, the integral is equal to the quadrature error, which is bounded by h^{p+1} times a bound of the p th derivative of the function $g(s)$. This derivative is bounded independently of h , because by Lemma 1.6 all derivatives of the collocation polynomial are bounded uniformly as $h \rightarrow 0$. Since, anyway, $p \leq 2s$, we get $y_1 - y(t_0 + h) = \mathcal{O}(h^{p+1})$ from (1.14). \square

Lemma 1.6. *The collocation polynomial $u(t)$ is an approximation of order s to the exact solution of (1.1) on the whole interval, i.e.,*

$$\|u(t) - y(t)\| \leq C \cdot h^{s+1} \quad \text{for } t \in [t_0, t_0 + h] \quad (1.15)$$

and for sufficiently small h .

Moreover, the derivatives of $u(t)$ satisfy for $t \in [t_0, t_0 + h]$

$$\|u^{(k)}(t) - y^{(k)}(t)\| \leq C \cdot h^{s+1-k} \quad \text{for } k = 0, \dots, s.$$

Proof. The collocation polynomial satisfies

$$\dot{u}(t_0 + \tau h) = \sum_{i=1}^s f(t_0 + c_i h, u(t_0 + c_i h)) \ell_i(\tau),$$

while the exact solution of (1.1) satisfies

$$\dot{y}(t_0 + \tau h) = \sum_{i=1}^s f(t_0 + c_i h, y(t_0 + c_i h)) \ell_i(\tau) + h^s E(\tau, h),$$

where the interpolation error $E(\tau, h)$ is bounded by $\max_{t \in [t_0, t_0 + h]} \|y^{(s+1)}(t)\|/s!$ and its derivatives satisfy

$$\|E^{(k-1)}(\tau, h)\| \leq \max_{t \in [t_0, t_0 + h]} \frac{\|y^{(s+1)}(t)\|}{(s - k + 1)!}.$$

This follows from the fact that, by Rolle's theorem, the differentiated polynomial $\sum_{i=1}^s f(t_0 + c_i h, y(t_0 + c_i h)) \ell_i^{(k-1)}(\tau)$ can be interpreted as the interpolation polynomial of $h^{k-1} y^{(k)}(t_0 + \tau h)$ at $s - k + 1$ points lying in $[t_0, t_0 + h]$. Integrating the difference of the above two equations gives

$$y(t_0 + \tau h) - u(t_0 + \tau h) = h \sum_{i=1}^s \Delta f_i \int_0^\tau \ell_i(\sigma) d\sigma + h^{s+1} \int_0^\tau E(\sigma, h) d\sigma \quad (1.16)$$

with $\Delta f_i = f(t_0 + c_i h, y(t_0 + c_i h)) - f(t_0 + c_i h, u(t_0 + c_i h))$. Using a Lipschitz condition for $f(t, y)$, this relation yields

$$\max_{t \in [t_0, t_0 + h]} \|y(t) - u(t)\| \leq h C L \max_{t \in [t_0, t_0 + h]} \|y(t) - u(t)\| + \text{Const} \cdot h^{s+1},$$

implying the statement (1.15) for sufficiently small $h > 0$.

The proof of the second statement follows from

$$h^k \left(y^{(k)}(t_0 + \tau h) - u^{(k)}(t_0 + \tau h) \right) = h \sum_{i=1}^s \Delta f_i \ell_i^{(k-1)}(\tau) + h^{s+1} E^{(k-1)}(\tau, h)$$

by using a Lipschitz condition for $f(t, y)$ and the estimate (1.15). \square

II.1.3 Gauss and Lobatto Collocation

Gauss Methods. If we take c_1, \dots, c_s as the zeros of the s th shifted Legendre polynomial

$$\frac{d^s}{dx^s} \left(x^s (x-1)^s \right),$$

the interpolatory quadrature formula has order $p = 2s$, and by Theorem 1.5, the Runge–Kutta (or collocation) method based on these nodes has the same order $2s$. For $s = 1$ we obtain the implicit midpoint rule. The Runge–Kutta coefficients for $s = 2$ (the method of Hammer & Hollingsworth 1955) and $s = 3$ are given in Table 1.1. The proof of the order properties for general s was a sensational result of Butcher (1964a). At that time these methods were considered, at least by the editors of *Math. of Comput.*, to be purely academic without any practical value; 5 years later their A -stability was discovered, 12 years later their B -stability, and 25 years later their symplecticity. Thus, of all the papers in issue No. 85 of *Math. of Comput.*, the one most important to us is the one for which publication was the most difficult.

Table 1.1. Gauss methods of order 4 and 6

$\frac{1}{2} - \frac{\sqrt{3}}{6}$	$\frac{1}{4}$	$\frac{1}{4} - \frac{\sqrt{3}}{6}$	
$\frac{1}{2} + \frac{\sqrt{3}}{6}$	$\frac{1}{4} + \frac{\sqrt{3}}{6}$	$\frac{1}{4}$	
	$\frac{1}{2}$	$\frac{1}{2}$	
$\frac{1}{2} - \frac{\sqrt{15}}{10}$	$\frac{5}{36}$	$\frac{2}{9} - \frac{\sqrt{15}}{15}$	$\frac{5}{36} - \frac{\sqrt{15}}{30}$
$\frac{1}{2}$	$\frac{5}{36} + \frac{\sqrt{15}}{24}$	$\frac{2}{9}$	$\frac{5}{36} - \frac{\sqrt{15}}{24}$
$\frac{1}{2} + \frac{\sqrt{15}}{10}$	$\frac{5}{36} + \frac{\sqrt{15}}{30}$	$\frac{2}{9} + \frac{\sqrt{15}}{15}$	$\frac{5}{36}$
	$\frac{5}{18}$	$\frac{4}{9}$	$\frac{5}{18}$

Radau Methods. Radau quadrature formulas have the highest possible order, $2s - 1$, among quadrature formulas with either $c_1 = 0$ or $c_s = 1$. The corresponding collocation methods for $c_s = 1$ are called Radau IIA methods. They play an important role in the integration of stiff differential equations (see Hairer & Wanner (1996), Sect. IV.8). However, they lack both *symmetry* and *symplecticity*, properties that will be the subjects of later chapters in this book.

Lobatto IIIA Methods. Lobatto quadrature formulas have the highest possible order with $c_1 = 0$ and $c_s = 1$. Under these conditions, the nodes must be the zeros of

$$\frac{d^{s-2}}{dx^{s-2}} \left(x^{s-1} (x-1)^{s-1} \right) \quad (1.17)$$

and the quadrature order is $p = 2s - 2$. The corresponding collocation methods are called, for historical reasons, Lobatto IIIA methods. For $s = 2$ we have the implicit trapezoidal rule. The coefficients for $s = 3$ and $s = 4$ are given in Table 1.2.

Table 1.2. Lobatto IIIA methods of order 4 and 6

0	0	0	0	0
$\frac{1}{2}$	$\frac{5}{24}$	$\frac{1}{3}$	$-\frac{1}{24}$	
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$	
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$	

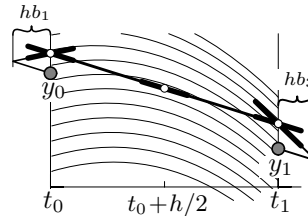
0	0	0	0	0
$\frac{5 - \sqrt{5}}{10}$	$\frac{11 + \sqrt{5}}{120}$	$\frac{25 - \sqrt{5}}{120}$	$\frac{25 - 13\sqrt{5}}{120}$	$\frac{-1 + \sqrt{5}}{120}$
$\frac{5 + \sqrt{5}}{10}$	$\frac{11 - \sqrt{5}}{120}$	$\frac{25 + 13\sqrt{5}}{120}$	$\frac{25 + \sqrt{5}}{120}$	$\frac{-1 - \sqrt{5}}{120}$
1	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$
	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$

II.1.4 Discontinuous Collocation Methods

Collocation methods allow, as we have seen above, a very elegant proof of their order properties. By similar ideas, they also admit strikingly simple proofs for their A - and B -stability as well as for symplecticity, our subject in Chap. VI. However, not all method classes are of collocation type. It is therefore interesting to define a modification of the collocation idea, which allows us to extend all the above proofs to much wider classes of methods. This definition will also lead, later, to important classes of *partitioned* methods.

Definition 1.7. Let c_2, \dots, c_{s-1} be distinct real numbers (usually $0 \leq c_i \leq 1$), and let b_1, b_s be two arbitrary real numbers. The corresponding *discontinuous collocation method* is then defined via a polynomial of degree $s - 2$ satisfying

$$\begin{aligned} u(t_0) &= y_0 - hb_1(\dot{u}(t_0) - f(t_0, u(t_0))) \\ \dot{u}(t_0 + c_i h) &= f(t_0 + c_i h, u(t_0 + c_i h)), \quad i = 2, \dots, s-1, \\ y_1 &= u(t_1) - hb_s(\dot{u}(t_1) - f(t_1, u(t_1))). \end{aligned} \quad (1.18)$$



The figure gives a geometric interpretation of the correction term in the first and third formulas of (1.18). The motivation for this definition will become clear in the proof of Theorem 1.9 below. Our first result shows that discontinuous collocation methods are equivalent to implicit Runge–Kutta methods.

Theorem 1.8. *The discontinuous collocation method of Definition 1.7 is equivalent to an s -stage Runge–Kutta method (1.4) with coefficients determined by $c_1 = 0$, $c_s = 1$, and*

$$\begin{aligned} a_{i1} &= b_1, & a_{is} &= 0 & \text{for } i &= 1, \dots, s, \\ C(s-2) & & \text{and} & & B(s-2), \end{aligned} \quad (1.19)$$

with the conditions $C(q)$ and $B(p)$ of (1.11).

Proof. As in the proof of Theorem 1.4 we put $k_i := \dot{u}(t_0 + c_i h)$ (this time for $i = 2, \dots, s-1$), so that $\dot{u}(t_0 + \tau h) = \sum_{j=2}^{s-1} k_j \cdot \ell_j(\tau)$ by the Lagrange interpolation formula. Here, $\ell_j(\tau)$ corresponds to c_2, \dots, c_{s-1} and is a polynomial of degree $s-3$. By integration and using the definition of $u(t_0)$ we get

$$\begin{aligned} u(t_0 + c_i h) &= u(t_0) + h \sum_{j=2}^{s-1} k_j \int_0^{c_i} \ell_j(\tau) d\tau \\ &= y_0 + h b_1 k_1 + h \sum_{j=2}^{s-1} k_j \left(\int_0^{c_i} \ell_j(\tau) d\tau - b_1 \ell_j(0) \right) \end{aligned}$$

with $k_1 = f(y_0)$. Inserted into (1.18) this gives the first formula of the Runge–Kutta equation (1.4) with $a_{ij} = \int_0^{c_i} \ell_j(\tau) d\tau - b_1 \ell_j(0)$. As for collocation methods, one checks that the a_{ij} are uniquely determined by the condition $C(s-2)$. The formula for y_1 is obtained similarly. \square

Table 1.3. Survey of discontinuous collocation methods

type	characteristics	prominent examples
$b_1 = 0, b_s = 0$	$(s-2)$ -stage collocation	Gauss, Radau IIA, Lobatto IIIA
$b_1 = 0, b_s \neq 0$	$(s-1)$ -stage with $a_{is} = 0$	methods of Butcher (1964b)
$b_1 \neq 0, b_s = 0$	$(s-1)$ -stage with $a_{i1} = b_1$	Radau IA, Lobatto IIIC
$b_1 \neq 0, b_s \neq 0$	s -stage with $a_{i1} = b_1, a_{is} = 0$	Lobatto IIIB

If $b_1 = 0$ in Definition 1.7, the entire first column in the Runge–Kutta tableau vanishes, so that the first stage can be removed, which leads to an equivalent method with $s-1$ stages. Similarly, if $b_s = 0$, we can remove the last stage. Therefore, we have all classes of methods, which are “continuous” either to the left, or to the right, or on both sides, as special cases in our definition.

In the case where $b_1 = b_s = 0$, the discontinuous collocation method (1.18) is equivalent to the $(s-2)$ -stage collocation method based on c_2, \dots, c_{s-1} (see Table 1.3). The methods with $b_s = 0$ but $b_1 \neq 0$, which include the Radau IA and

Table 1.4. Lobatto IIIB methods of order 4 and 6

				0	$\frac{1}{12}$	$\frac{-1-\sqrt{5}}{24}$	$\frac{-1+\sqrt{5}}{24}$	0
				$\frac{5-\sqrt{5}}{10}$	$\frac{1}{12}$	$\frac{25+\sqrt{5}}{120}$	$\frac{25-13\sqrt{5}}{120}$	0
				$\frac{5+\sqrt{5}}{10}$	$\frac{1}{12}$	$\frac{25+13\sqrt{5}}{120}$	$\frac{25-\sqrt{5}}{120}$	0
				1	$\frac{1}{12}$	$\frac{11-\sqrt{5}}{24}$	$\frac{11+\sqrt{5}}{24}$	0
0	$\frac{1}{6}$	$-\frac{1}{6}$	0		$\frac{1}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{1}{12}$
$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{3}$	0					
1	$\frac{1}{6}$	$\frac{5}{6}$	0					

Lobatto IIIC methods, are of interest for the solution of stiff differential equations (Hairer & Wanner 1996). The methods with $b_1 = 0$ but $b_s \neq 0$, introduced by Butcher (1964a, 1964b), are of historical interest. They were thought to be computationally attractive, because their last stage is explicit. In the context of geometric integration, much more important are methods for which both $b_1 \neq 0$ and $b_s \neq 0$.

Lobatto IIIB Methods (Table 1.4). We consider the quadrature formulas whose nodes are the zeros of (1.17). We have $c_1 = 0$ and $c_s = 1$. Based on c_2, \dots, c_{s-1} and b_1, b_s we consider the discontinuous collocation method. This class of methods is called Lobatto IIIB (Ehle 1969), and it plays an important role in geometric integration in conjunction with the Lobatto IIIA methods of Sect. II.1.3 (see Theorem IV.2.3 and Theorem VI.4.5). These methods are of order $2s-2$, as the following result shows.

Theorem 1.9 (Superconvergence). *The discontinuous collocation method of Definition 1.7 has the same order as the underlying quadrature formula.*

Proof. We follow the lines of the proof of Theorem 1.5. With the polynomial $u(t)$ of Definition 1.7, and with the defect

$$\delta(t) := \dot{u}(t) - f(t, u(t))$$

we get (1.13) after linearization. The variation of constants formula then yields

$$\begin{aligned} u(t_0 + h) - y(t_0 + h) &= R(t_0 + h, t_0)(u(t_0) - y_0) \\ &+ \int_{t_0}^{t_0+h} R(t_0 + h, s) \left(\delta(s) + r(s) \right) ds, \end{aligned}$$

which corresponds to (1.14) if $u(t_0) = y_0$. As a consequence of Lemma 1.10 below (with $k = 0$), the integral over $R(t_0 + h, s)r(s)$ gives a $\mathcal{O}(h^{2s-1})$ contribution. Since the defect $\delta(t_0 + c_i h)$ vanishes only for $i = 2, \dots, s-1$, an application of the quadrature formula to $R(t_0 + h, s)\delta(s)$ yields $hb_1 R(t_0 + h, t_0)\delta(t_0) + hb_s \delta(t_0 + h)$ in addition to the quadrature error, which is $\mathcal{O}(h^{p+1})$. Collecting terms suitably, we obtain

$$u(t_1) - hb_s \delta(t_1) - y(t_1) = R(t_1, t_0)(u(t_0) + hb_1 \delta(t_0) - y_0) + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{2s-1}),$$

which, after using the definitions of $u(t_0)$ and $u(t_1)$, proves $y_1 - y(t_1) = \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{2s-1})$. \square

Lemma 1.10. *The polynomial $u(t)$ of the discontinuous collocation method (1.18) satisfies for $t \in [t_0, t_0 + h]$ and for sufficiently small h*

$$\|u^{(k)}(t) - y^{(k)}(t)\| \leq C \cdot h^{s-1-k} \quad \text{for } k = 0, \dots, s-2.$$

Proof. The proof is essentially the same as that for Lemma 1.6. In the formulas for $\dot{u}(t_0 + \tau h)$ and $\dot{y}(t_0 + \tau h)$, the sum has to be taken from $i = 2$ to $i = s-1$. Moreover, all h^s become h^{s-2} . In (1.16) one has an additional term

$$y_0 - u(t_0) = hb_1(\dot{u}(t_0) - f(t_0, u(t_0))),$$

which, however, is just an interpolation error of size $\mathcal{O}(h^{s-1})$ and can be included in $\text{Const} \cdot h^{s-1}$. \square

II.2 Partitioned Runge–Kutta Methods

Some interesting numerical methods introduced in Chap. I (symplectic Euler and the Störmer–Verlet method) do not belong to the class of Runge–Kutta methods. They are important examples of so-called partitioned Runge–Kutta methods. In this section we consider differential equations in the partitioned form

$$\dot{y} = f(y, z), \quad \dot{z} = g(y, z), \quad (2.1)$$

where y and z may be vectors of different dimensions.

II.2.1 Definition and First Examples

The idea is to take two different Runge–Kutta methods, and to treat the y -variables with the first method (a_{ij}, b_i) , and the z -variables with the second method $(\hat{a}_{ij}, \hat{b}_i)$.

Definition 2.1. Let b_i, a_{ij} and \hat{b}_i, \hat{a}_{ij} be the coefficients of two Runge–Kutta methods. A *partitioned Runge–Kutta method* for the solution of (2.1) is given by

$$\begin{aligned} k_i &= f\left(y_0 + h \sum_{j=1}^s a_{ij} k_j, z_0 + h \sum_{j=1}^s \hat{a}_{ij} \ell_j\right), \\ \ell_i &= g\left(y_0 + h \sum_{j=1}^s a_{ij} k_j, z_0 + h \sum_{j=1}^s \hat{a}_{ij} \ell_j\right), \\ y_1 &= y_0 + h \sum_{i=1}^s b_i k_i, \quad z_1 = z_0 + h \sum_{i=1}^s \hat{b}_i \ell_i. \end{aligned} \quad (2.2)$$

Methods of this type were originally proposed by Hofer in 1976 and by Gripenberg in 1978 for problems with stiff and nonstiff parts (see Hairer, Nørsett & Wanner (1993), Sect. II.15). Their importance for Hamiltonian systems (see the examples of Chap. I) has been discovered only in the last decade.

An interesting example is the symplectic Euler method (I.1.9), where the implicit Euler method $b_1 = 1, a_{11} = 1$ is combined with the explicit Euler method $\hat{b}_1 = 1, \hat{a}_{11} = 0$. The Störmer–Verlet method (I.1.17) is of the form (2.2) with coefficients given in Table 2.1.

Table 2.1. Störmer–Verlet as a partitioned Runge–Kutta method

0	0	0	1/2	1/2	0
1	1/2	1/2	1/2	1/2	0
	1/2	1/2		1/2	1/2

The theory of Runge–Kutta methods can be extended in a straightforward manner to partitioned methods. Since (2.2) is a one-step method $(y_1, z_1) = \Phi_h(y_0, z_0)$, the Definition 1.2 of the order applies directly. Considering problems $\dot{y} = f(y)$, $\dot{z} = g(z)$ without any coupling terms, we see that the order of (2.2) cannot exceed $\min(p, \hat{p})$, where p and \hat{p} are the orders of the two methods.

Conditions for Order Two. Expanding the exact solution of (2.1) and the numerical solution (2.2) into Taylor series, we see that the method is of order 2 if the coupling conditions

$$\sum_{ij} b_i \hat{a}_{ij} = 1/2, \quad \sum_{ij} \hat{b}_i a_{ij} = 1/2 \quad (2.3)$$

are satisfied in addition to the usual Runge–Kutta order conditions for order 2. The method of Table 2.1 satisfies these conditions, and it is therefore of order 2. We also remark that (2.3) is automatically satisfied by partitioned methods that are based on the same quadrature nodes, i.e.,

$$c_i = \hat{c}_i \quad \text{for all } i \quad (2.4)$$

where, as usual, $c_i = \sum_j a_{ij}$ and $\hat{c}_i = \sum_j \hat{a}_{ij}$.

Conditions for Order Three. The conditions for order three already become quite complicated, unless (2.4) is satisfied. In this case, we obtain the additional conditions

$$\sum_{ij} b_i \hat{a}_{ij} c_j = 1/6, \quad \sum_{ij} \hat{b}_i a_{ij} c_j = 1/6. \quad (2.5)$$

The order conditions for higher order will be discussed in Sect. III.2.2. It turns out that the number of coupling conditions increases very fast with order, and the proofs for high order are often very cumbersome. There is, however, a very elegant proof of the order for the partitioned method which is the most important one in connection with “geometric integration”, as we shall see now.

II.2.2 Lobatto IIIA–IIIB Pairs

These methods generalize the Störmer–Verlet method to arbitrary order. Indeed, the left method of Table 2.1 is the trapezoidal rule, which is the Lobatto IIIA method with $s = 2$, and the method to the right is equivalent to the midpoint rule and, apart from the values of the c_i , is the Lobatto IIIB method with $s = 2$. Sun (1993b) and Jay (1996) discovered that for general s the combination of the Lobatto IIIA and IIIB methods are suitable for Hamiltonian systems. The coefficients of the methods for $s = 3$ are given in Table 2.2. Using the idea of discontinuous collocation, we give a direct proof of the order for this pair of methods.

Table 2.2. Coefficients of the 3-stage Lobatto IIIA–IIIB pair

0	0	0	0	0	1/6	-1/6	0
1/2	5/24	1/3	-1/24	1/2	1/6	1/3	0
1	1/6	2/3	1/6	1	1/6	5/6	0
	1/6	2/3	1/6		1/6	2/3	1/6

Theorem 2.2. *The partitioned Runge–Kutta method composed of the s -stage Lobatto IIIA and the s -stage Lobatto IIIB method, is of order $2s - 2$.*

Proof. Let $c_1 = 0, c_2, \dots, c_{s-1}, c_s = 1$ and b_1, \dots, b_s be the nodes and weights of the Lobatto quadrature. The partitioned Runge–Kutta method based on the Lobatto IIIA–IIIB pair can be interpreted as the discontinuous collocation method

$$\begin{aligned}
 u(t_0) &= y_0 \\
 v(t_0) &= z_0 - hb_1(\dot{v}(t_0) - g(u(t_0), v(t_0))) \\
 \dot{u}(t_0 + c_i h) &= f(u(t_0 + c_i h), v(t_0 + c_i h)), & i = 1, \dots, s \\
 \dot{v}(t_0 + c_i h) &= g(u(t_0 + c_i h), v(t_0 + c_i h)), & i = 2, \dots, s-1 \\
 y_1 &= u(t_1) \\
 z_1 &= v(t_1) - hb_s(\dot{v}(t_1) - g(u(t_1), v(t_1))),
 \end{aligned} \tag{2.6}$$

where $u(t)$ and $v(t)$ are polynomials of degree s and $s-2$, respectively. This is seen as in the proofs of Theorem 1.4 and Theorem 1.8. The superconvergence (order $2s - 2$) is obtained with exactly the same proof as for Theorem 1.9, where the functions $u(t)$ and $y(t)$ have to be replaced with $(u(t), v(t))^T$ and $(y(t), z(t))^T$, etc. Instead of Lemma 1.10 we use the estimates (for $t \in [t_0, t_0 + h]$)

$$\begin{aligned}
 \|u^{(k)}(t) - y^{(k)}(t)\| &\leq c \cdot h^{s-k} \quad \text{for } k = 0, \dots, s, \\
 \|v^{(k)}(t) - z^{(k)}(t)\| &\leq c \cdot h^{s-1-k} \quad \text{for } k = 0, \dots, s-2,
 \end{aligned}$$

which can be proved by following the lines of the proofs of Lemma 1.6 and Lemma 1.10. \square

II.2.3 Nyström Methods

Da bis jetzt die *direkte* Anwendung der Rungeschen Methode auf den wichtigen Fall von Differentialgleichungen zweiter Ordnung nicht behandelt war ... (E.J. Nyström 1925)

Second-order differential equations

$$\ddot{y} = g(t, y, \dot{y}) \quad (2.7)$$

form an important class of problems. Most of the differential equations in Chap. I are of this form (e.g., the Kepler problem, the outer solar system, problems in molecular dynamics). This is mainly due to Newton's law that forces are proportional to second derivatives (acceleration). Introducing a new variable $z = \dot{y}$ for the first derivative, the problem (2.7) becomes equivalent to the partitioned system

$$\dot{y} = z, \quad \dot{z} = g(t, y, z). \quad (2.8)$$

A partitioned Runge–Kutta method (2.2) applied to this system yields

$$\begin{aligned} k_i &= z_0 + h \sum_{j=1}^s \hat{a}_{ij} \ell_j, \\ \ell_i &= g\left(t_0 + c_i h, y_0 + h \sum_{j=1}^s a_{ij} k_j, z_0 + h \sum_{j=1}^s \hat{a}_{ij} \ell_j\right), \\ y_1 &= y_0 + h \sum_{i=1}^s b_i k_i, \quad z_1 = z_0 + h \sum_{i=1}^s \hat{b}_i \ell_i. \end{aligned} \quad (2.9)$$

If we insert the formula for k_i into the others, we obtain Definition 2.3 with

$$\bar{a}_{ij} = \sum_{k=1}^s a_{ik} \hat{a}_{kj}, \quad \bar{b}_i = \sum_{k=1}^s b_k \hat{a}_{ki}. \quad (2.10)$$

Definition 2.3. Let $c_i, \bar{b}_i, \bar{a}_{ij}$ and \hat{b}_i, \hat{a}_{ij} be real coefficients. A *Nyström method* for the solution of (2.7) is given by

$$\begin{aligned} \ell_i &= g\left(t_0 + c_i h, y_0 + c_i h \dot{y}_0 + h^2 \sum_{j=1}^s \bar{a}_{ij} \ell_j, \dot{y}_0 + h \sum_{j=1}^s \hat{a}_{ij} \ell_j\right), \\ y_1 &= y_0 + h \dot{y}_0 + h^2 \sum_{i=1}^s \bar{b}_i \ell_i, \quad \dot{y}_1 = \dot{y}_0 + h \sum_{i=1}^s \hat{b}_i \ell_i. \end{aligned} \quad (2.11)$$

For the important special case $\ddot{y} = g(t, y)$, where the vector field does not depend on the velocity, the coefficients \hat{a}_{ij} need not be specified. A Nyström method is of order p if $y_1 - y(t_0 + h) = \mathcal{O}(h^{p+1})$ and $\dot{y}_1 - \dot{y}(t_0 + h) = \mathcal{O}(h^{p+1})$. It is not sufficient to consider y_1 alone. The order conditions will be discussed in Sect. III.2.3.

Notice that the Störmer–Verlet scheme (I.1.17) is a Nyström method for problems of the form $\ddot{y} = g(t, y)$. We have $s = 2$, and the coefficients are $c_1 = 0, c_2 = 1, \bar{a}_{11} = \bar{a}_{12} = \bar{a}_{22} = 0, \bar{a}_{21} = 1/2, \bar{b}_1 = 1/2, \bar{b}_2 = 0$, and $\hat{b}_1 = \hat{b}_2 = 1/2$. With $q_{n+1/2} = q_n + \frac{h}{2} v_{n+1/2}$ the step $(q_{n-1/2}, v_{n-1/2}) \mapsto (q_{n+1/2}, v_{n+1/2})$ of (I.1.17) becomes a one-stage Nyström method with $c_1 = 1/2, \bar{a}_{11} = 0, \bar{b}_1 = \hat{b}_1 = 1$.

II.3 The Adjoint of a Method

We shall see in Chap. V that *symmetric* numerical methods have many important properties. The key for understanding symmetry is the concept of the *adjoint* method.

The flow φ_t of an autonomous differential equation

$$\dot{y} = f(y), \quad y(t_0) = y_0 \quad (3.1)$$

satisfies $\varphi_{-t}^{-1} = \varphi_t$. This property is *not*, in general, shared by the one-step map Φ_h of a numerical method. An illustration is presented in the upper picture of Fig. 3.1 (a), where we see that the one-step map Φ_h for the explicit Euler method is different from the inverse of Φ_{-h} , which is the implicit Euler method.

Definition 3.1. The *adjoint method* Φ_h^* of a method Φ_h is the inverse map of the original method with reversed time step $-h$, i.e.,

$$\Phi_h^* := \Phi_{-h}^{-1} \quad (3.2)$$

(see Fig. 3.1 (b)). In other words, $y_1 = \Phi_h^*(y_0)$ is implicitly defined by $\Phi_{-h}(y_1) = y_0$. A method for which $\Phi_h^* = \Phi_h$ is called *symmetric*.

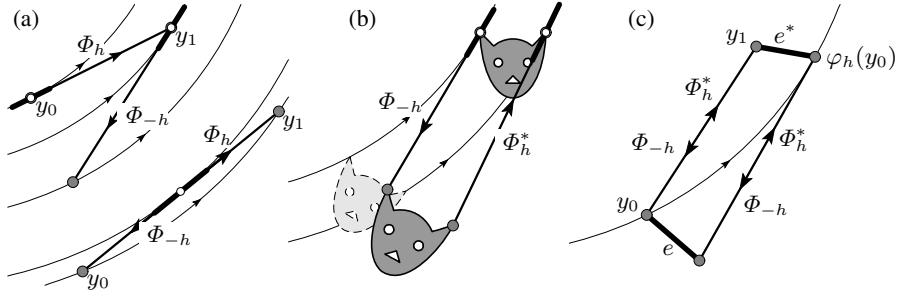


Fig. 3.1. Definition and properties of the adjoint method

The consideration of adjoint methods evolved independently from the study of symmetric integrators (Stetter (1973), p. 125, Wanner (1973)) and from the aim of constructing and analyzing stiff integrators from explicit ones (Cash (1975) calls them “the backward version” which were the first example of mono-implicit methods and Scherer (1977) calls them “reflected methods”).

The adjoint method satisfies the usual properties such as $(\Phi_h^*)^* = \Phi_h$ and $(\Phi_h \circ \Psi_h)^* = \Psi_h^* \circ \Phi_h^*$ for any two one-step methods Φ_h and Ψ_h . The implicit Euler method is the adjoint of the explicit Euler method. The implicit midpoint rule is symmetric (see the lower picture of Fig. 3.1 (a)), and the trapezoidal rule and the Störmer–Verlet method are also symmetric.

The following theorem shows that the adjoint method has the same order as the original method, and, with a possible sign change, also the same leading error term.

Theorem 3.2. Let φ_t be the exact flow of (3.1) and let Φ_h be a one-step method of order p satisfying

$$\Phi_h(y_0) = \varphi_h(y_0) + C(y_0)h^{p+1} + \mathcal{O}(h^{p+2}). \quad (3.3)$$

The adjoint method Φ_h^* then has the same order p and we have

$$\Phi_h^*(y_0) = \varphi_h(y_0) + (-1)^p C(y_0)h^{p+1} + \mathcal{O}(h^{p+2}). \quad (3.4)$$

If the method is symmetric, its (maximal) order is even.

Proof. The idea of the proof is exhibited in drawing (c) of Fig. 3.1. From a given initial value y_0 we compute $\varphi_h(y_0)$ and $y_1 = \Phi_h^*(y_0)$, whose difference e^* is the local error of Φ_h^* . This error is then “projected back” by Φ_{-h} to become e . We see that $-e$ is the local error of Φ_{-h} , i.e., by hypothesis (3.3),

$$e = (-1)^p C(\varphi_h(y_0))h^{p+1} + \mathcal{O}(h^{p+2}). \quad (3.5)$$

Since $\varphi_h(y_0) = y_0 + \mathcal{O}(h)$ and $e = (I + \mathcal{O}(h))e^*$, it follows that

$$e^* = (-1)^p C(y_0)h^{p+1} + \mathcal{O}(h^{p+2})$$

which proves (3.4). The statement for symmetric methods is an immediate consequence of this result, because $\Phi_h = \Phi_h^*$ implies $C(y_0) = (-1)^p C(y_0)$, and therefore $C(y_0)$ can be different from zero only for even p . \square

II.4 Composition Methods

The idea of composing methods has some tradition in several variants: composition of different Runge–Kutta methods with the same step size leading to the Butcher group, which is treated in Sect. III.1.3; cyclic composition of multistep methods for breaking the “Dahlquist barrier” (see Stetter (1973), p. 216); composition of low order Runge–Kutta methods for increasing stability for stiff problems (Gentzsch & Schlüter (1978), Iserles (1984)). In the following, we consider the composition of a given basic one-step method (and, eventually, its adjoint method) with *different* step sizes. The aim is to increase the order while preserving some desirable properties of the basic method. This idea has mainly been developed in the papers of Suzuki (1990), Yoshida (1990), and McLachlan (1995).

Let Φ_h be a basic method and $\gamma_1, \dots, \gamma_s$ real numbers. Then we call its composition with step sizes $\gamma_1 h, \gamma_2 h, \dots, \gamma_s h$, i.e.,

$$\Psi_h = \Phi_{\gamma_s h} \circ \dots \circ \Phi_{\gamma_1 h}, \quad (4.1)$$

the corresponding *composition method* (see Fig. 4.1 (a)).

Theorem 4.1. *Let Φ_h be a one-step method of order p . If*

$$\begin{aligned} \gamma_1 + \dots + \gamma_s &= 1 \\ \gamma_1^{p+1} + \dots + \gamma_s^{p+1} &= 0, \end{aligned} \quad (4.2)$$

then the composition method (4.1) is at least of order $p + 1$.

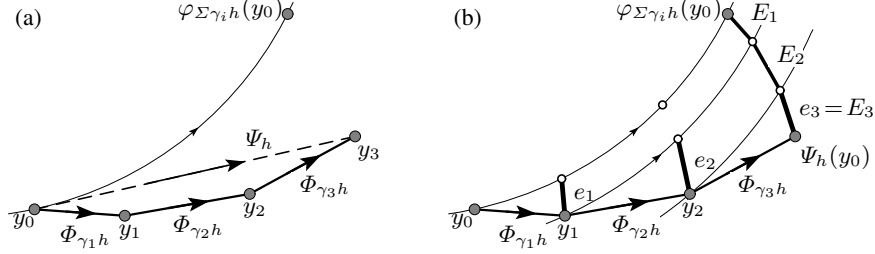


Fig. 4.1. Composition of method Φ_h with three step sizes

Proof. The proof is presented in Fig. 4.1 (b) for $s = 3$. It is very similar to the proof of Theorem 3.2. By hypothesis

$$\begin{aligned} e_1 &= C(y_0) \cdot \gamma_1^{p+1} h^{p+1} + \mathcal{O}(h^{p+2}) \\ e_2 &= C(y_1) \cdot \gamma_2^{p+1} h^{p+1} + \mathcal{O}(h^{p+2}) \\ e_3 &= C(y_2) \cdot \gamma_3^{p+1} h^{p+1} + \mathcal{O}(h^{p+2}). \end{aligned} \quad (4.3)$$

We have, as before, $y_i = y_0 + \mathcal{O}(h)$ and $E_i = (I + \mathcal{O}(h))e_i$ for all i and obtain, for $\sum \gamma_i = 1$,

$$\varphi_h(y_0) - \Psi_h(y_0) = E_1 + E_2 + E_3 = C(y_0)(\gamma_1^{p+1} + \gamma_2^{p+1} + \gamma_3^{p+1})h^{p+1} + \mathcal{O}(h^{p+2})$$

which shows that under conditions (4.2) the $\mathcal{O}(h^{p+1})$ -term vanishes. \square

Example 4.2 (The Triple Jump). Equations (4.2) have no real solution for odd p . Therefore, the order increase is only possible for even p . In this case, the smallest s which allows a solution is $s = 3$. We then have some freedom for solving the two equations. If we impose symmetry $\gamma_1 = \gamma_3$, then we obtain (Creutz & Gocksch 1989, Forest 1989, Suzuki 1990, Yoshida 1990)

$$\gamma_1 = \gamma_3 = \frac{1}{2 - 2^{1/(p+1)}}, \quad \gamma_2 = -\frac{2^{1/(p+1)}}{2 - 2^{1/(p+1)}}. \quad (4.4)$$

This procedure can be repeated: we start with a symmetric method of order 2, apply (4.4) with $p = 2$ to obtain order 3; due to the symmetry of the γ 's this new method is in fact of order 4 (see Theorem 3.2). With this new method we repeat (4.4) with $p = 4$ and obtain a symmetric 9-stage composition method of order 6, then with $p = 6$ a 27-stage symmetric composition method of order 8, and so on. One obtains in this way *any* order, however, at the price of a terrible zig-zag of the step points (see Fig. 4.2).

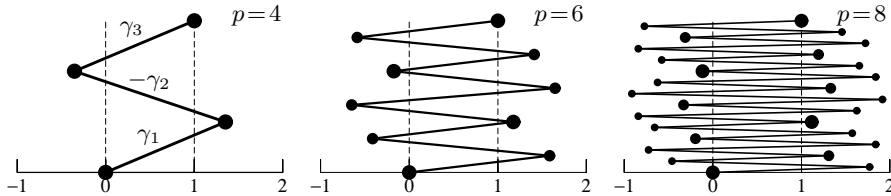


Fig. 4.2. The Triple Jump of order 4 and its iterates of orders 6 and 8

Example 4.3 (Suzuki's Fractals). If one desires methods with smaller values of γ_i , one has to increase s even more. For example, for $s = 5$ the best solution of (4.2) has the sign structure $++-++$ with $\gamma_1 = \gamma_2$ (see Exercise 7). This leads to (Suzuki 1990)

$$\gamma_1 = \gamma_2 = \gamma_4 = \gamma_5 = \frac{1}{4 - 4^{1/(p+1)}}, \quad \gamma_3 = -\frac{4^{1/(p+1)}}{4 - 4^{1/(p+1)}}. \quad (4.5)$$

The repetition of this algorithm for $p = 2, 4, 6, \dots$ leads to a fractal structure of the step points (see Fig. 4.3).

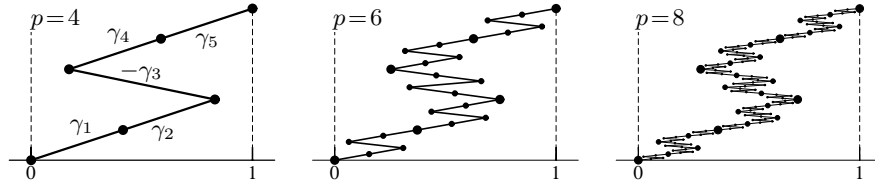


Fig. 4.3. Suzuki's "fractal" composition methods

Composition with the Adjoint Method. If we replace the composition (4.1) by the more general formula

$$\Psi_h = \Phi_{\alpha_s h} \circ \Phi_{\beta_s h}^* \circ \dots \circ \Phi_{\beta_2 h}^* \circ \Phi_{\alpha_1 h} \circ \Phi_{\beta_1 h}^*, \quad (4.6)$$

the condition for order $p+1$ becomes, by using the result (3.4) and a similar proof as above,

$$\begin{aligned} \beta_1 + \alpha_1 + \beta_2 + \dots + \beta_s + \alpha_s &= 1 \\ (-1)^p \beta_1^{p+1} + \alpha_1^{p+1} + (-1)^p \beta_2^{p+1} + \dots + (-1)^p \beta_s^{p+1} + \alpha_s^{p+1} &= 0. \end{aligned} \quad (4.7)$$

This allows an order increase for odd p as well. In particular, we see at once the solution $\alpha_1 = \beta_1 = 1/2$ for $p = s = 1$, which turns every consistent one-step method of order 1 into a second-order symmetric method

$$\Psi_h = \Phi_{h/2} \circ \Phi_{h/2}^*. \quad (4.8)$$

Example 4.4. If Φ_h is the explicit (resp. implicit) Euler method, then Ψ_h in (4.8) becomes the implicit midpoint (resp. trapezoidal) rule.

Example 4.5. In a second-order problem $\dot{q} = p$, $\dot{p} = g(q)$, if Φ_h is the symplectic Euler method, which discretizes q by the implicit Euler and p by the explicit Euler method, then the composed method Ψ_h in (4.8) is the Störmer–Verlet method (I.1.17).

A Numerical Example. To demonstrate the numerical performance of the above methods, we choose the Kepler problem (I.2.2) with $e = 0.6$ and the initial values from (I.2.11). As integration interval we choose $[0, 7.5]$, a bit more than one revolution. The exact solution is obtained by carefully evaluating the integral (I.2.10), which gives

$$\varphi = 8.67002632314281495159108828552, \quad (4.9)$$

with the help of which we compute r , $\dot{\varphi}$, \dot{r} from (I.2.8) and (I.2.6). This gives

$$\begin{aligned} q_1 &= -0.828164402690770818204757585370 \\ q_2 &= 0.778898095658635447081654480796 \\ p_1 &= -0.856384715343395351524486215030 \\ p_2 &= -0.160552150799838435254419104102. \end{aligned} \quad (4.10)$$

As the basic method we use the Verlet scheme and compare in Fig. 4.4 the performances of the composition sequences of the Triple Jump (4.4) and those of Suzuki (4.5) for a large number of different equidistant basic step sizes and for orders $p = 4, 6, 8, 10, 12$. Each basic step is then divided into 3, 9, 27, 81, 243 respectively 5, 25, 125, 625, 3125 composition steps and the maximal final error is compared with the total number of function evaluations in double logarithmic scales. For each method and order, all the points lie asymptotically on a straight line with slope $-p$. Therefore, theoretically, a higher order method will become superior when the precision requirements become sufficiently high. But we see that for orders 10 and 12 these “break even points” are far beyond any precision of practical interest, after some 40 or 50 digits. We also observe that the wild zig-zag of the Triple Jump (4.4) is a more serious handicap than the enormous number of small steps of the Suzuki sequence (4.5).

For later reference we have also included, in black symbols, the results obtained by the two methods (V.3.11) and (V.3.13) of orders 6 and 8, respectively, which will be the outcome of a more elaborate order theory of Chap. III.

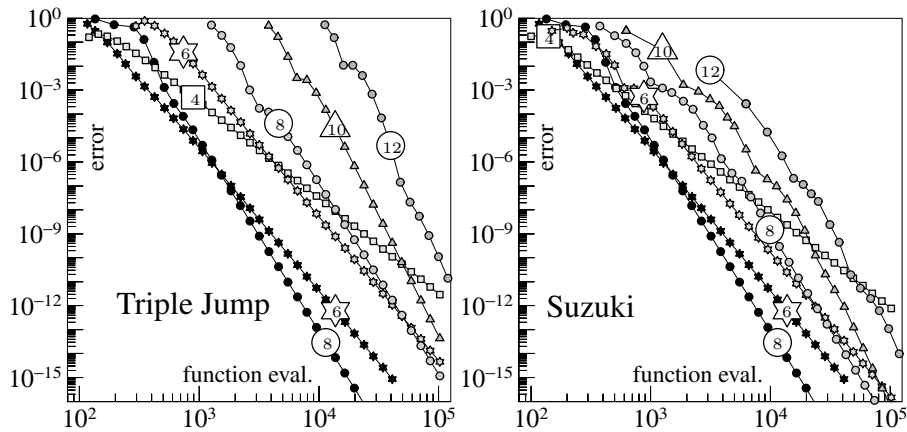


Fig. 4.4. Numerical results of the Triple Jump and Suzuki step sequences (grey symbols) compared to optimal methods (black symbols)

II.5 Splitting Methods

The splitting idea yields an approach that is completely different from Runge–Kutta methods. One decomposes the vector field into integrable pieces and treats them separately.

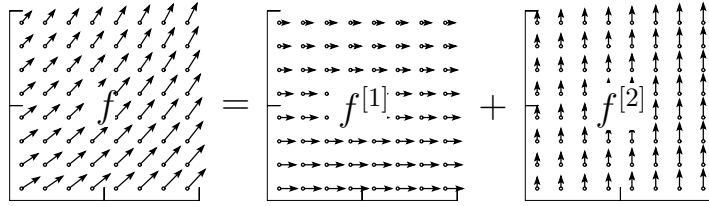


Fig. 5.1. A splitting of a vector field

We consider an arbitrary system $\dot{y} = f(y)$ in \mathbb{R}^n , and suppose that the vector field is “split” as (see Fig. 5.1)

$$\dot{y} = f^{[1]}(y) + f^{[2]}(y). \quad (5.1)$$

If then, by chance, the exact flows $\varphi_t^{[1]}$ and $\varphi_t^{[2]}$ of the systems $\dot{y} = f^{[1]}(y)$ and $\dot{y} = f^{[2]}(y)$ can be calculated explicitly, we can, from a given initial value y_0 , first solve the first system to obtain a value $y_{1/2}$, and from this value integrate the second system to obtain y_1 . In this way we have introduced the numerical methods

$$\begin{aligned} \Phi_h^* &= \varphi_h^{[2]} \circ \varphi_h^{[1]} \\ \Phi_h &= \varphi_h^{[1]} \circ \varphi_h^{[2]} \end{aligned} \quad \begin{array}{c} \text{Diagram 1: } y_0 \xrightarrow{\varphi_h^{[1]}} y_{1/2} \xrightarrow{\varphi_h^{[2]}} y_1 \\ \text{Diagram 2: } y_0 \xrightarrow{\varphi_h^{[2]}} y_{1/2} \xrightarrow{\varphi_h^{[1]}} y_1 \end{array} \quad (5.2)$$

where one is the adjoint of the other. These formulas are often called the *Lie–Trotter splitting* (Trotter 1959). By Taylor expansion we find that $(\varphi_h^{[1]} \circ \varphi_h^{[2]})(y_0) = \varphi_h(y_0) + \mathcal{O}(h^2)$, so that both methods give approximations of order 1 to the solution of (5.1). Another idea is to use a symmetric version and put

$$\Phi_h^{[S]} = \varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}, \quad \begin{array}{c} \text{Diagram: } y_0 \xrightarrow{\varphi_{h/2}^{[1]}} y_{1/2} \xrightarrow{\varphi_h^{[2]}} y_1 \end{array} \quad (5.3)$$

which is known as the *Strang splitting*¹ (Strang 1968), and sometimes as the *Marchuk splitting* (Marchuk 1968). By breaking up in (5.3) $\varphi_h^{[2]} = \varphi_{h/2}^{[2]} \circ \varphi_{h/2}^{[2]}$,

¹ The article Strang (1968) deals with spatial discretizations of partial differential equations such as $u_t = Au_x + Bu_y$. There, the functions $f^{[i]}$ typically contain differences in only one spatial direction.

we see that the Strang splitting $\Phi_h^{[S]} = \Phi_{h/2} \circ \Phi_{h/2}^*$ is the composition of the Lie-Trotter method and its adjoint with halved step sizes. The Strang splitting formula is therefore symmetric and of order 2 (see (4.8)).

Example 5.1 (The Symplectic Euler and the Störmer–Verlet Schemes). Suppose we have a Hamiltonian system with separable Hamiltonian $H(p, q) = T(p) + U(q)$. We consider this as the sum of *two* Hamiltonians, the first one depending only on p , the second one only on q . The corresponding Hamiltonian systems

$$\begin{aligned} \dot{p} &= 0 & \text{and} & & \dot{p} &= -U_q(q) \\ \dot{q} &= T_p(p) & & & \dot{q} &= 0 \end{aligned} \quad (5.4)$$

can be solved without problem to yield

$$\begin{aligned} p(t) &= p_0 & \text{and} & & p(t) &= p_0 - t U_q(q_0) \\ q(t) &= q_0 + t T_p(p_0) & & & q(t) &= q_0. \end{aligned} \quad (5.5)$$

Denoting the flows of these two systems by φ_t^T and φ_t^U , we see that the symplectic Euler method (I.1.9) is just the composition $\varphi_h^T \circ \varphi_h^U$. Furthermore, the adjoint of the symplectic Euler method is $\varphi_h^U \circ \varphi_h^T$, and by Example 4.5 the Verlet scheme is $\varphi_{h/2}^U \circ \varphi_h^T \circ \varphi_{h/2}^U$, the Strang splitting (5.3). Anticipating the results of Chap. VI, the flows φ_h^T and φ_h^U are both symplectic transformations, and, since the composition of symplectic maps is again symplectic, this gives an elegant proof of the symplecticity of the “symplectic” Euler method and the Verlet scheme.

General Splitting Procedure. In a similar way to the general idea of composition methods (4.6), we can form with arbitrary coefficients $a_1, b_1, a_2, \dots, a_m, b_m$ (where, eventually, a_1 or b_m , or both, are zero)

$$\Psi_h = \varphi_{b_m h}^{[2]} \circ \varphi_{a_m h}^{[1]} \circ \varphi_{b_{m-1} h}^{[2]} \circ \dots \circ \varphi_{a_2 h}^{[1]} \circ \varphi_{b_1 h}^{[2]} \circ \varphi_{a_1 h}^{[1]} \quad (5.6)$$

and try to increase the order of the scheme by suitably determining the free coefficients. An early contribution to this subject is the article of Ruth (1983), where, for the special case (5.4), a method (5.6) of order 3 with $m = 3$ is constructed. Forest & Ruth (1990) and Candy & Rozmus (1991) extend Ruth’s technique and construct methods of order 4. One of their methods is just (4.1) with $\gamma_1, \gamma_2, \gamma_3$ given by (4.4) ($p = 2$) and Φ_h from (5.3). A systematic study of such methods started with the articles of Suzuki (1990, 1992) and Yoshida (1990).

A close connection between the theories of splitting methods (5.6) and of composition methods (4.6) was discovered by McLachlan (1995). Indeed, if we put $\beta_1 = a_1$ and break up $\varphi_{b_1 h}^{[2]} = \varphi_{\alpha_1 h}^{[2]} \circ \varphi_{\beta_1 h}^{[2]}$ (group property of the exact flow) where α_1 is given in (5.8), further $\varphi_{a_2 h}^{[1]} = \varphi_{\beta_2 h}^{[1]} \circ \varphi_{\alpha_1 h}^{[1]}$ and so on (cf. Fig. 5.2), we see, using (5.2), that Ψ_h of (5.6) is identical with Ψ_h of (4.6), where

$$\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]} \quad \text{so that} \quad \Phi_h^* = \varphi_h^{[2]} \circ \varphi_h^{[1]}. \quad (5.7)$$

A necessary and sufficient condition for the existence of α_i and β_i satisfying (5.8) is that $\sum a_i = \sum b_i$, which is the consistency condition anyway for method (5.6).

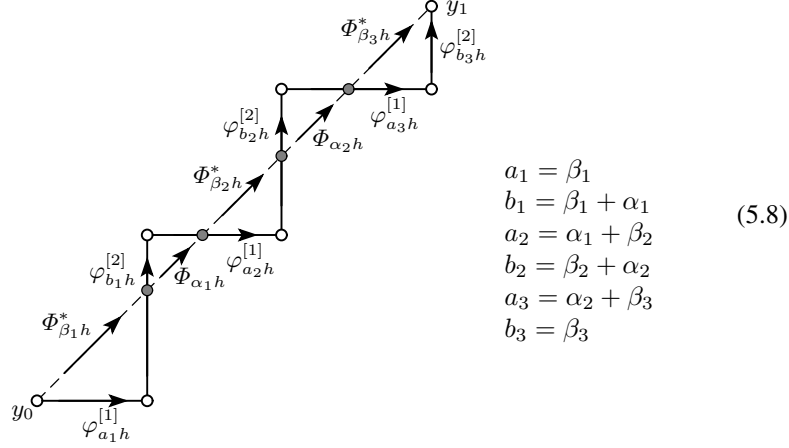


Fig. 5.2. Equivalence of splitting and composition methods

Combining Exact and Numerical Flows. It may happen that the differential equation $\dot{y} = f(y)$ can be split according to (5.1), such that only the flow of, say, $\dot{y} = f^{[1]}(y)$ can be computed exactly. If $f^{[1]}(y)$ constitutes the dominant part of the vector field, it is natural to search for integrators that exploit this information. The above interpretation of splitting methods as composition methods allows us to construct such integrators. We just consider

$$\Phi_h = \varphi_h^{[1]} \circ \Phi_h^{[2]}, \quad \Phi_h^* = \Phi_h^{[2]*} \circ \varphi_h^{[1]} \quad (5.9)$$

as the basis of the composition method (4.6). Here $\varphi_t^{[1]}$ is the exact flow of $\dot{y} = f^{[1]}(y)$, and $\Phi_h^{[2]}$ is some first-order integrator applied to $\dot{y} = f^{[2]}(y)$. Since Φ_h of (5.9) is consistent with (5.1), the resulting method (4.6) has the desired high order. It is given by

$$\Psi_h = \varphi_{\alpha_s h}^{[1]} \circ \Phi_{\alpha_s h}^{[2]} \circ \Phi_{\beta_s h}^{[2]*} \circ \varphi_{(\beta_s + \alpha_{s-1})h}^{[1]} \circ \Phi_{\alpha_{s-1} h}^{[2]} \circ \dots \circ \Phi_{\beta_1 h}^{[2]*} \circ \varphi_{\beta_1 h}^{[1]}. \quad (5.10)$$

Notice that replacing $\varphi_t^{[2]}$ with a low-order approximation $\Phi_t^{[2]}$ in (5.6) would not retain the high order of the composition, because $\Phi_t^{[2]}$ does not satisfy the group property.

Splitting into More than Two Vector Fields. Consider a differential equation

$$\dot{y} = f^{[1]}(y) + f^{[2]}(y) + \dots + f^{[N]}(y), \quad (5.11)$$

where we assume that the flows $\varphi_t^{[j]}$ of the individual problems $\dot{y} = f^{[j]}(y)$ can be computed exactly. In this case there are many possibilities for extending (5.6) and for writing the method as a composition of $\varphi_{a_j h}^{[1]}, \varphi_{b_j h}^{[2]}, \varphi_{c_j h}^{[3]}, \dots$. This makes it difficult to find optimal compositions of high order. A simple and efficient way is to consider the first-order method

$$\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]} \circ \dots \circ \varphi_h^{[N]}$$

together with its adjoint as the basis of the composition (4.6). Without any additional effort this yields splitting methods for (5.11) of arbitrary high order.

II.6 Exercises

1. Compute all collocation methods with $s = 2$ as a function of c_1 and c_2 . Which of them are of order 3, which of order 4?
2. Prove that the collocation solution plotted in the right picture of Fig. 1.3 is composed of arcs of parabolas.
3. Let $b_1 = b_4 = 1/8$, $c_2 = 1/3$, $c_3 = 2/3$, and consider the corresponding discontinuous collocation method. Determine its order and find the coefficients of the equivalent Runge–Kutta method.
4. Show that each of the symplectic Euler methods in (I.1.9) is the adjoint of the other.
5. (Additive Runge–Kutta methods). Let b_i, a_{ij} and \hat{b}_i, \hat{a}_{ij} be the coefficients of two Runge–Kutta methods. An additive Runge–Kutta method for the solution of $\dot{y} = f^{[1]}(y) + f^{[2]}(y)$ is given by

$$\begin{aligned} k_i &= f^{[1]}\left(y_0 + h \sum_{j=1}^s a_{ij} k_j\right) + f^{[2]}\left(y_0 + h \sum_{j=1}^s \hat{a}_{ij} k_j\right) \\ y_1 &= y_0 + h \sum_{i=1}^s b_i k_i. \end{aligned}$$

Show that this can be interpreted as a partitioned Runge–Kutta method (2.2) applied to

$$\dot{y} = f^{[1]}(y) + f^{[2]}(z), \quad \dot{z} = f^{[1]}(y) + f^{[2]}(z)$$

with $y(0) = z(0) = y_0$. Notice that $y(t) = z(t)$.

6. Let Φ_h denote the Störmer–Verlet scheme, and consider the composition

$$\Phi_{\gamma_{2k+1}h} \circ \Phi_{\gamma_{2k}h} \circ \dots \circ \Phi_{\gamma_2h} \circ \Phi_{\gamma_1h}$$

with $\gamma_1 = \dots = \gamma_k = \gamma_{k+2} = \dots = \gamma_{2k+1}$. Compute γ_1 and γ_{k+1} such that the composition gives a method of order 4. For several differential equations (pendulum, Kepler problem) study the global error of a constant step size implementation as a function of k .

7. Consider the composition method (4.1) with $s = 5$, $\gamma_5 = \gamma_1$, and $\gamma_4 = \gamma_2$. Among the solutions of

$$2\gamma_1 + 2\gamma_2 + \gamma_3 = 1, \quad 2\gamma_1^3 + 2\gamma_2^3 + \gamma_3^3 = 0$$

find the one that minimizes $|2\gamma_1^5 + 2\gamma_2^5 + \gamma_3^5|$.

Remark. This property motivates the choice of the γ_i in (4.5).

Chapter III.

Order Conditions, Trees and B-Series

In this chapter we present a compact theory of the order conditions of the methods presented in Chap. II, in particular Runge–Kutta methods, partitioned Runge–Kutta methods, and composition methods by using the notion of rooted trees and B-series. These ideas lead to algebraic structures which have recently found interesting applications in quantum field theory. The chapter terminates with the Baker–Campbell–Hausdorff formula, which allows another access to the order properties of composition and splitting methods.

Some parts of this chapter are rather short, but nevertheless self-contained. For more detailed presentations we refer to the monographs of Butcher (1987), of Hairer, Nørsett & Wanner (1993), and of Hairer & Wanner (1996). Readers mainly interested in geometric properties of numerical integrators may continue with Chapters IV, V or VI before returning to the technically more difficult jungle of trees.

III.1 Runge–Kutta Order Conditions and B-Series

Even the standard notation has been found to be too heavy in dealing with
fourth and higher order processes, . . . (R.H. Merson 1957)

In this section we derive the order conditions of Runge–Kutta methods by comparing the Taylor series of the exact solution of (1.1) with that of the numerical solution. The computation is much simplified, first by considering an *autonomous* system of equations (Gill 1951), and second, by the use of rooted trees (connected graphs without cycles and a distinguished vertex; Merson 1957). The theory has been developed by Butcher in the years 1963–72 (see Butcher (1987), Sect. 30) and by Hairer & Wanner in 1973–74 (see Hairer, Nørsett & Wanner (1993), Sections II.2 and II.12). Here we give new simplified proofs.

III.1.1 Derivation of the Order Conditions

We consider an autonomous problem

$$\dot{y} = f(y), \quad y(t_0) = y_0, \quad (1.1)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is sufficiently differentiable. A problem $\dot{y} = f(t, y)$ can be brought into this form by appending the equation $\dot{t} = 1$. We develop the subsequent theory in four steps.

Er sagte es klar und angenehm,
was erstens, zweitens und drittens käm'. (W. Busch, *Jobsiade* 1872)

First Step. We compute the higher derivatives of the solution y at the initial point t_0 . For this, we have from (1.1)

$$y^{(q)} = (f(y))^{(q-1)} \quad (1.2)$$

and compute the latter derivatives by using the chain rule, the product rule, the symmetry of partial derivatives, and the notation $f'(y)$ for the derivative as a linear map (the Jacobian), $f''(y)$ the second derivative as a bilinear map and similarly for higher derivatives. This gives

$$\begin{aligned} \dot{y} &= f(y) \\ \ddot{y} &= f'(y) \dot{y} \\ y^{(3)} &= f''(y)(\dot{y}, \dot{y}) + f'(y) \ddot{y} \\ y^{(4)} &= f'''(y)(\dot{y}, \dot{y}, \dot{y}) + 3f''(y)(\ddot{y}, \dot{y}) + f'(y) y^{(3)} \\ y^{(5)} &= f^{(4)}(y)(\dot{y}, \dot{y}, \dot{y}, \dot{y}) + 6f'''(y)(\ddot{y}, \dot{y}, \dot{y}) + 4f''(y)(y^{(3)}, \dot{y}) \\ &\quad + 3f''(y)(\ddot{y}, \ddot{y}) + f'(y) y^{(4)}, \end{aligned} \quad (1.3)$$

and so on. The coefficients 3, 6, 4, 3, ... appearing in these expressions have a certain combinatorial meaning (number of partitions of a set of $q-1$ elements), but for the moment we need not know their values.

Second Step. We insert in (1.3) recursively the computed derivatives \dot{y}, \ddot{y}, \dots into the right side of the subsequent formulas. This gives for the first few

$$\begin{aligned} \dot{y} &= f \\ \ddot{y} &= f'f \\ y^{(3)} &= f''(f, f) + f'f'f \\ y^{(4)} &= f'''(f, f, f) + 3f''(f'f, f) + f'f''(f, f) + f'f'f'f, \end{aligned} \quad (1.4)$$

where the arguments (y) have been suppressed. The expressions which appear in these formulas, denoted by $F(\tau)$, will be called the *elementary differentials*. We represent each of them by a suitable graph τ (a rooted tree) as follows:

Each f becomes a vertex, a first derivative f' becomes a vertex with one branch, and a k th derivative $f^{(k)}$ becomes a vertex with k branches pointing upwards. The arguments of the k -linear mapping $f^{(k)}(y)$ correspond to trees that are attached on the upper ends of these branches. The tree to the right corresponds to $f''(f'f, f)$. Other trees are plotted in Table 1.1. In the above process, each insertion of an already known derivative consists of grafting the corresponding trees upon a new root as in Definition 1.1 below, and inserting the corresponding elementary differentials as arguments of $f^{(m)}(y)$ as in Definition 1.2.

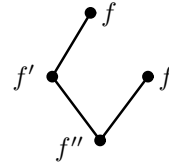










Table 1.1. Trees, elementary differentials, and coefficients

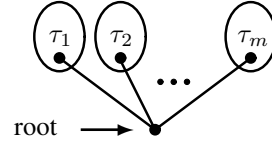
$ \tau $	τ	graph	$\alpha(\tau)$	$F(\tau)$	$\gamma(\tau)$	$\phi(\tau)$	$\sigma(\tau)$
1	\bullet		1	f	1	$\sum_i b_i$	1
2	$[\bullet]$		1	$f'f$	2	$\sum_{ij} b_i a_{ij}$	1
3	$[\bullet, \bullet]$		1	$f''(f, f)$	3	$\sum_{ijk} b_i a_{ij} a_{ik}$	2
3	$[[\bullet]]$		1	$f'f'f$	6	$\sum_{ijk} b_i a_{ij} a_{jk}$	1
4	$[\bullet, \bullet, \bullet]$		1	$f'''(f, f, f)$	4	$\sum_{ijkl} b_i a_{ij} a_{ik} a_{il}$	6
4	$[[\bullet], \bullet]$		3	$f''(f'f, f)$	8	$\sum_{ijkl} b_i a_{ij} a_{ik} a_{jl}$	1
4	$[[\bullet, \bullet]]$		1	$f'f''(f, f)$	12	$\sum_{ijkl} b_i a_{ij} a_{jk} a_{jl}$	2
4	$[[[\bullet]]]$		1	$f'f'f'f$	24	$\sum_{ijkl} b_i a_{ij} a_{jk} a_{kl}$	1

Definition 1.1 (Trees). The set of (rooted) *trees* T is recursively defined as follows:

- the graph \bullet with only one vertex (called the root) belongs to T ;
- if $\tau_1, \dots, \tau_m \in T$, then the graph obtained by grafting the roots of τ_1, \dots, τ_m to a new vertex also belongs to T . It is denoted by

$$\tau = [\tau_1, \dots, \tau_m],$$

and the new vertex is the root of τ .



We further denote by $|\tau|$ the *order* of τ (the number of vertices), and by $\alpha(\tau)$ the coefficients appearing in the formulas (1.4). We remark that some of the trees among τ_1, \dots, τ_m may be equal and that τ does not depend on the ordering of τ_1, \dots, τ_m . For example, we do not distinguish between $[[\bullet], \bullet]$ and $[\bullet, [\bullet]]$.

Definition 1.2 (Elementary Differentials). For a tree $\tau \in T$ the *elementary differential* is a mapping $F(\tau) : \mathbb{R}^n \rightarrow \mathbb{R}^n$, defined recursively by $F(\bullet)(y) = f(y)$ and

$$F(\tau)(y) = f^{(m)}(y) \left(F(\tau_1)(y), \dots, F(\tau_m)(y) \right) \quad \text{for } \tau = [\tau_1, \dots, \tau_m].$$

Examples of these constructions and the corresponding coefficients are seen in Table 1.1. With these definitions, we obtain from (1.4):

Theorem 1.3. The q th derivative of the exact solution is given by

$$y^{(q)}(t_0) = \sum_{|\tau|=q} \alpha(\tau) F(\tau)(y_0), \quad (1.5)$$

where $\alpha(\tau)$ are positive integer coefficients. \square

Third Step. We now turn to the numerical solution of the Runge–Kutta method (II.1.4), which, by putting $hk_i = g_i$, we write as

$$g_i = hf(u_i) \quad (1.6)$$

and

$$u_i = y_0 + \sum_j a_{ij} g_j, \quad y_1 = y_0 + \sum_i b_i g_i, \quad (1.7)$$

where u_i , g_i and y_1 are functions of h . We develop the derivatives of (1.6), by Leibniz' rule, and obtain $g_i^{(q)} = h(f(u_i))^{(q)} + q \cdot (f(u_i))^{(q-1)}$. This gives, for $h = 0$,

$$g_i^{(q)} = q \cdot (f(u_i))^{(q-1)}, \quad (1.8)$$

the same expression as in (1.2), with y just replaced by u_i and with an extra factor q . Consequently, exactly as in (1.3),

$$\begin{aligned} \dot{g}_i &= 1 \cdot f(y_0) \\ \ddot{g}_i &= 2 \cdot f'(y_0) \dot{u}_i \\ g_i^{(3)} &= 3 \cdot (f''(y_0)(\dot{u}_i, \dot{u}_i) + f'(y_0) \ddot{u}_i) \\ g_i^{(4)} &= 4 \cdot (f'''(y_0)(\dot{u}_i, \dot{u}_i, \dot{u}_i) + 3f''(y_0)(\ddot{u}_i, \dot{u}_i) + f'(y_0) u_i^{(3)}) \\ g_i^{(5)} &= 5 \cdot (f^{(4)}(y_0)(\dot{u}_i, \dot{u}_i, \dot{u}_i, \dot{u}_i) + 6f'''(y_0)(\ddot{u}_i, \dot{u}_i, \dot{u}_i) + 4f''(y_0)(u_i^{(3)}, \dot{u}_i) \\ &\quad + 3f''(y_0)(\ddot{u}_i, \ddot{u}_i) + f'(y_0) u_i^{(4)}), \end{aligned} \quad (1.9)$$

and so on. Here, the derivatives of g_i and u_i are evaluated at $h = 0$.

Fourth Step. We now insert recursively the derivatives $\dot{u}_i, \ddot{u}_i, \dots$ into (1.9). This will give the next higher derivative of g_i , and, using

$$u_i^{(q)} = \sum_j a_{ij} \cdot g_j^{(q)}, \quad (1.10)$$

which follows from (1.7), also the next higher derivative of u_i . This process begins as

$$\begin{aligned} \dot{g}_i &= 1 \cdot f & \dot{u}_i &= 1 \cdot (\sum_j a_{ij}) \cdot f \\ \ddot{g}_i &= (1 \cdot 2) (\sum_j a_{ij}) f' f & \ddot{u}_i &= (1 \cdot 2) (\sum_{jk} a_{ij} a_{jk}) f' f \end{aligned} \quad (1.11)$$

and so on. If we compare these formulas with the first lines of (1.4), we see that the results are precisely the same, apart from the extra factors. We denote the *integer factors* $1, 1 \cdot 2, \dots$ by $\gamma(\tau)$ and the factors containing the a_{ij} 's by $\mathbf{g}_i(\tau)$ and $\mathbf{u}_i(\tau)$, respectively. We obtain by induction that the same happens in general, i.e. that, in contrast to (1.5),

$$\begin{aligned}
g_i^{(q)}|_{h=0} &= \sum_{|\tau|=q} \gamma(\tau) \cdot \mathbf{g}_i(\tau) \cdot \alpha(\tau) F(\tau)(y_0) \\
u_i^{(q)}|_{h=0} &= \sum_{|\tau|=q} \gamma(\tau) \cdot \mathbf{u}_i(\tau) \cdot \alpha(\tau) F(\tau)(y_0),
\end{aligned} \tag{1.12}$$

where $\alpha(\tau)$ and $F(\tau)$ are *the same* quantities as before. This is seen by continuing the insertion process of the derivatives $u_i^{(q)}$ into the right-hand side of (1.9). For example, if \dot{u}_i and \ddot{u}_i are inserted into $3f''(\ddot{u}_i, \dot{u}_i)$, we will obtain the corresponding expression as in (1.4), multiplied by the two extra factors $\mathbf{u}_i(\text{J})$, brought in by \ddot{u}_i , and $\mathbf{u}_i(\bullet)$ from \dot{u}_i . For a general tree $\tau = [\tau_1, \dots, \tau_m]$ this will be

$$\mathbf{g}_i(\tau) = \mathbf{u}_i(\tau_1) \cdot \dots \cdot \mathbf{u}_i(\tau_m). \tag{1.13}$$

Second, the factors $\gamma(\text{J})$ and $\gamma(\bullet)$ will receive the additional factor $q = |\tau|$ from (1.9), i.e., we will have in general

$$\gamma(\tau) = |\tau| \gamma(\tau_1) \cdot \dots \cdot \gamma(\tau_m). \tag{1.14}$$

Then, by (1.10),

$$\mathbf{u}_i(\tau) = \sum_j a_{ij} \mathbf{g}_j(\tau) = \sum_j a_{ij} \cdot \mathbf{u}_j(\tau_1) \cdot \dots \cdot \mathbf{u}_j(\tau_m). \tag{1.15}$$

This formula can be re-used repeatedly, as long as some of the trees τ_1, \dots, τ_m are of order > 1 . Finally, we have from the last formula of (1.7), that the coefficients for the numerical solution, which we denote by $\phi(\tau)$ and call the *elementary weights*, satisfy

$$\phi(\tau) = \sum_i b_i \mathbf{g}_i(\tau). \tag{1.16}$$

We summarize the result as follows:

Theorem 1.4. *The derivatives of the numerical solution of a Runge–Kutta method (II.1.4), for $h = 0$, are given by*

$$y_1^{(q)}|_{h=0} = \sum_{|\tau|=q} \gamma(\tau) \cdot \phi(\tau) \cdot \alpha(\tau) F(\tau)(y_0), \tag{1.17}$$

where $\alpha(\tau)$ and $F(\tau)$ are the same as in Theorem 1.3, the coefficients $\gamma(\tau)$ satisfy $\gamma(\bullet) = 1$ and (1.14). The elementary weights $\phi(\tau)$ are obtained from the tree τ as follows: attach to every vertex a summation letter (“ i ” to the root), then $\phi(\tau)$ is the sum, over all summation indices, of a product composed of b_i , and factors a_{jk} for each vertex “ j ” directly connected with “ k ” by an upwards directed branch.

Proof. Repeated application of (1.15) followed by (1.16) shows that the elementary weight $\phi(\tau)$ is the collection of $\sum_i b_i$ from (1.16) and all $\sum_j a_{ij}$ of (1.15). \square

Theorem 1.5. *The Runge–Kutta method has order p if and only if*

$$\phi(\tau) = \frac{1}{\gamma(\tau)} \quad \text{for } |\tau| \leq p. \quad (1.18)$$

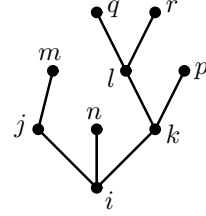
Proof. The comparison of Theorem 1.3 with Theorem 1.4 proves the sufficiency of condition (1.18). The necessity of (1.18) follows from the independence of the elementary differentials (see e.g., Hairer, Nørsett & Wanner (1993), Exercise 4 of Sect. II.2). \square

Example 1.6. For the following tree of order 9 we have

$$\sum_{i,j,k,l,m,n,p,q,r} b_i a_{ij} a_{jm} a_{in} a_{ik} a_{kl} a_{lq} a_{lr} a_{kp} = \frac{1}{9 \cdot 2 \cdot 5 \cdot 3}$$

or, by using $\sum_j a_{ij} = c_i$,

$$\sum_{i,j,k,l} b_i c_i a_{ij} c_j a_{ik} c_k a_{kl} c_l^2 = \frac{1}{270}.$$



The quantities $\phi(\tau)$ and $\gamma(\tau)$ for all trees up to order 4 are given in Table 1.1. This also verifies the formulas (II.1.6) stated previously.

III.1.2 B-Series

We now introduce the concept of B-series, which gives further insight into the behaviour of numerical methods and allows extensions to more general classes of methods.

Motivated by formulas (1.12) and (1.17) above, we consider the corresponding *series* as the objects of our study. This means, we study power series in $h^{|\tau|}$ containing elementary differentials $F(\tau)$ and arbitrary coefficients which are now written in the form $a(\tau)$. Such series will be called B-series. To move from (1.6) to (1.13) we need to prove a result stating that *a B-series inserted into $hf(\cdot)$ is again a B-series*. We start with

$$B(a, y) = y + a(\bullet)hf(y) + a(\text{J})h^2(f'f)(y) + \dots = y + \delta, \quad (1.19)$$

and get by Taylor expansion

$$hf(B(a, y)) = hf(y + \delta) = hf(y) + hf'(y)\delta + \frac{h}{2!}f''(y)(\delta, \delta) + \dots \quad (1.20)$$

Inserting δ from (1.19) and multiplying out, we obtain the expression

$$\begin{aligned} hf(B(a, y)) &= hf + h^2 a(\bullet) f'f + h^3 a(\text{J}) f' f' f + \frac{h^3}{2!} a(\bullet)^2 f''(f, f) \\ &\quad + h^4 a(\bullet) a(\text{J}) f''(f'f, f) + \dots \end{aligned} \quad (1.21)$$

This beautiful formula is not yet perfect for two reasons. First, there is a denominator $2!$ in the fourth term. The origin of this lies in the *symmetry* of the tree \mathbf{V} . We thus introduce the symmetry coefficients of Definition 1.7 (following Butcher 1987, Theorem 144A). Second, there is no first term y . We therefore allow the factor $a(\emptyset)$ in Definition 1.8.

Definition 1.7 (Symmetry coefficients). The symmetry coefficients $\sigma(\tau)$ are defined by $\sigma(\bullet) = 1$ and, for $\tau = [\tau_1, \dots, \tau_m]$,

$$\sigma(\tau) = \sigma(\tau_1) \cdot \dots \cdot \sigma(\tau_m) \cdot \mu_1! \mu_2! \cdot \dots, \quad (1.22)$$

where the integers μ_1, μ_2, \dots count equal trees among τ_1, \dots, τ_m .

Definition 1.8 (B-Series). For a mapping $a : T \cup \{\emptyset\} \rightarrow \mathbb{R}$ a formal series of the form

$$B(a, y) = a(\emptyset)y + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau) F(\tau)(y) \quad (1.23)$$

is called a *B-series*.¹

The main results of the theory of B-series have their origin in the paper of Butcher (1972), although series expansions were not used there. B-series were then introduced by Hairer & Wanner (1974). The normalization used in Definition 1.8 is due to Butcher & Sanz-Serna (1996). The following fundamental lemma gives a second way of finding the order conditions.

Lemma 1.9. *Let $a : T \cup \{\emptyset\} \rightarrow \mathbb{R}$ be a mapping satisfying $a(\emptyset) = 1$. Then the corresponding B-series inserted into $hf(\cdot)$ is again a B-series. That is*

$$hf(B(a, y)) = B(a', y), \quad (1.24)$$

where $a'(\emptyset) = 0$, $a'(\bullet) = 1$, and

$$a'(\tau) = a(\tau_1) \cdot \dots \cdot a(\tau_m) \quad \text{for } \tau = [\tau_1, \dots, \tau_m]. \quad (1.25)$$

Proof. Since $a(\emptyset) = 1$ we have $B(a, y) = y + \mathcal{O}(h)$, so that $hf(B(a, y))$ can be expanded into a Taylor series around y . As in formulas (1.20) and (1.21), we get

¹ In this section we are not concerned about the convergence of the series. We shall see later in Chap. IX that the series converges for sufficiently small h , if $a(\tau)$ satisfies an inequality $|a(\tau)| \leq \gamma(\tau)cd^{|\tau|}$ and if $f(y)$ is an analytic function. If $f(y)$ is only k -times differentiable, then all formulas of this section remain valid for the truncated B-series $\sum_{\tau \in T, |\tau| \leq k} \cdot / \cdot$ with a suitable remainder term of size $\mathcal{O}(h^{k+1})$ added.

$$\begin{aligned}
hf(B(a, y)) &= h \sum_{m \geq 0} \frac{1}{m!} f^{(m)}(y) (B(a, y) - y)^m \\
&= h \sum_{m \geq 0} \frac{1}{m!} \sum_{\tau_1 \in T} \cdots \sum_{\tau_m \in T} \frac{h^{|\tau_1| + \dots + |\tau_m|}}{\sigma(\tau_1) \cdots \sigma(\tau_m)} \cdot a(\tau_1) \cdots a(\tau_m) \\
&\quad \cdot f^{(m)}(y) (F(\tau_1)(y), \dots, F(\tau_m)(y)) \\
&= \sum_{m \geq 0} \sum_{\tau_1 \in T} \cdots \sum_{\tau_m \in T} \frac{h^{|\tau|}}{\sigma(\tau)} \frac{\mu_1! \mu_2! \cdots}{m!} \cdot a'(\tau) F(\tau)(y) \\
&\quad \text{with } \tau = [\tau_1, \dots, \tau_m] \\
&= \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} a'(\tau) F(\tau)(y) = B(a', y).
\end{aligned}$$

The last equality follows from the fact that there are $\binom{m}{\mu_1, \mu_2, \dots}$ possibilities for writing the tree τ in the form $\tau = [\tau_1, \dots, \tau_m]$. For example, the trees $[\bullet, \bullet, [\bullet]]$, $[\bullet, [\bullet], \bullet]$ and $[[\bullet], \bullet, \bullet]$ appear as different terms in the upper sum, but only as one term in the lower sum. \square

Back to the Order Conditions. We present now a new derivation of the order conditions that is solely based on B-series and on Lemma 1.9. Let a Runge–Kutta method, say formulas (1.6) and (1.7), be given. All quantities in the defining formulas are set up as B-series, $g_i = B(\mathbf{g}_i, y_0)$, $u_i = B(\mathbf{u}_i, y_0)$, $y_1 = B(\phi, y_0)$. Then, either the linearity and/or Lemma 1.9, translate the formulas of the method into corresponding formulas for the coefficients (1.13), (1.15), and (1.16). This recursively justifies the ansatz as B-series.

Assuming the *exact* solution to be a B-series $B(\mathbf{e}, y_0)$, a term-by-term derivation of this series and an application of Lemma 1.9 to (1.1) yields

$$\mathbf{e}(\tau) = \frac{1}{|\tau|} \mathbf{e}(\tau_1) \cdots \mathbf{e}(\tau_m).$$

Together with definition (1.14) of $\gamma(\tau)$ we thus obtain

$$\mathbf{e}(\tau) = \frac{1}{\gamma(\tau)}. \quad (1.26)$$

A comparison of the coefficients of the B-series $y_1 = B(\phi, y_0)$ with those of the exact solution gives (1.18) and proves Theorem 1.5 again.

Comparing the B-series $B(\mathbf{e}, y_0)$ for the exact solution with Theorem 1.3, we get as a byproduct the formula

$$\alpha(\tau) = \frac{|\tau|!}{\sigma(\tau) \cdot \gamma(\tau)}. \quad (1.27)$$

If the available tools are enriched by the more general composition law of Theorem 1.10 below, this procedure can be applied to yet larger classes of methods.

III.1.3 Composition of Methods

The order theory for the composition of methods goes back to 1969, when Butcher used it to circumvent the order barrier for explicit 5th order 5 stage methods. It led to the seminal publication of Butcher (1972), where the general composition formula in (1.34) was expressed recursively.

Composition of Runge–Kutta Methods. Suppose that, starting from an initial value y_0 , we compute a numerical solution y_1 using a Runge–Kutta method with coefficients a_{ij}, b_i and step size h . Then, continuing from y_1 , we compute a value y_2 using another method with coefficients a_{ij}^*, b_i^* and the same step size. This composition of two methods is now considered as a *single* method (with coefficients \hat{a}_{ij}, \hat{b}_i). The problem is to derive the order properties of this new method, in particular to express the elementary weights $\hat{\phi}(\tau)$ in terms of those of the original two methods.

If the value y_1 from the first method is inserted into the starting value for the second method, one sees that the coefficients of the combined method are given by (here written for two-stage methods)

$$\begin{array}{c|c} \hat{a}_{11} & \hat{a}_{12} \\ \hat{a}_{21} & \hat{a}_{22} \\ \hat{a}_{31} & \hat{a}_{32} & \hat{a}_{33} & \hat{a}_{34} \\ \hat{a}_{41} & \hat{a}_{42} & \hat{a}_{43} & \hat{a}_{44} \\ \hline \hat{b}_1 & \hat{b}_2 & \hat{b}_3 & \hat{b}_4 \end{array} = \begin{array}{c|c|c|c} a_{11} & a_{12} & & \\ a_{21} & a_{22} & & \\ b_1 & b_2 & a_{11}^* & a_{12}^* \\ b_1 & b_2 & a_{21}^* & a_{22}^* \\ \hline b_1 & b_2 & b_1^* & b_2^* \end{array} \quad (1.28)$$

and our problem is to compute the elementary weights of this scheme.

Derivation. The idea is to write the sum for $\hat{\phi}(\tau)$, say for the tree $\hat{\mathcal{V}}$, in full detail

$$\hat{\phi}(\hat{\mathcal{V}}) = \sum_{i=1}^4 \sum_{j=1}^4 \sum_{k=1}^4 \sum_{l=1}^4 \hat{b}_i \hat{a}_{ij} \hat{a}_{ik} \hat{a}_{kl} = \dots \quad (1.29)$$

and to split each sum into the two different index sets. This leads to $2^{|\tau|}$ different expressions $\sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 \sum_{l=1}^2 \cdot / \cdot + \sum_{i=3}^4 \sum_{j=1}^2 \sum_{k=1}^2 \sum_{l=1}^2 \cdot / \cdot + \sum_{i=1}^2 \sum_{j=3}^4 \sum_{k=1}^2 \sum_{l=1}^2 \cdot / \cdot + \dots$. We symbolize each expression by drawing the corresponding vertex of τ as a *bullet* for the first index set and as a *star* for the second. However, due to the zero pattern in the matrix in (1.28) (the upper right corner is missing), each term with “star above bullet” can be omitted, since the corresponding \hat{a}_{ij} ’s are zero. So the only combinations to be considered are those of Fig. 1.1. We finally insert the quantities from the right tableau in (1.28),

$$\begin{aligned} \hat{\phi}(\hat{\mathcal{V}}) = & \sum b_i a_{ij} a_{ik} a_{kl} + \sum b_i^* b_j b_k a_{kl} + \sum b_i^* a_{ij}^* b_k a_{kl} + \sum b_i^* b_j a_{ik}^* b_l \\ & + \sum b_i^* a_{ij}^* a_{ik}^* b_l + \sum b_i^* b_j a_{ik}^* a_{kl} + \sum b_i^* a_{ij}^* a_{ik}^* a_{kl}, \end{aligned}$$

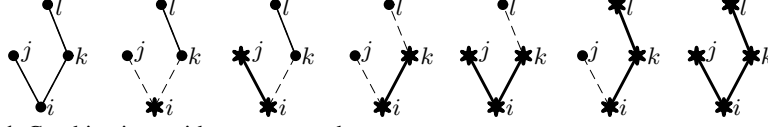


Fig. 1.1. Combinations with nonzero product

and we observe that each factor of the type b_j interrupts the summation, so that the terms decompose into factors of elementary weights of the individual methods as follows:

$$\begin{aligned} \widehat{\phi}(\text{tree}) &= \phi(\text{tree}) + \phi^*(\bullet) \cdot \phi(\bullet) \phi(\text{tree}) + \phi^*(\text{tree}) \cdot \phi(\text{tree}) + \phi^*(\text{tree}) \cdot \phi(\bullet) \phi(\bullet) \\ &\quad + \phi^*(\text{tree}) \cdot \phi(\bullet) + \phi^*(\text{tree}) \cdot \phi(\bullet) + \phi^*(\text{tree}) . \end{aligned}$$

The trees composed of the “star” nodes of τ in Fig. 1.1 constitute all possible “sub-trees” θ (from the empty tree to τ itself) having the same root as τ . This is the key for understanding the general result.

Ordered Trees. In order to formalize the procedure of Fig. 1.1, we introduce the set OT of *ordered trees* recursively as follows: $\bullet \in OT$, and

$$\text{if } \omega_1, \dots, \omega_m \in OT, \text{ then also the ordered } m\text{-tuple } (\omega_1, \dots, \omega_m) \in OT. \quad (1.30)$$

As the name suggests, in the graphical representation of an ordered tree the order of the branches leaving cannot be permuted. Neglecting the ordering, a tree $\tau \in T$ can be considered as an equivalence class of ordered trees, denoted $\tau = \overline{\omega}$.

For example, the tree of Fig. 1.1 has two orderings, namely tree_1 and tree_2 . We denote by $\nu(\tau)$ the number of possible orderings of the tree τ . It is given by $\nu(\bullet) = 1$ and

$$\nu(\tau) = \frac{m!}{\mu_1! \mu_2! \dots} \nu(\tau_1) \cdot \dots \cdot \nu(\tau_m) \quad (1.31)$$

for $\tau = [\tau_1, \dots, \tau_m]$, where the integers μ_1, μ_2, \dots are the numbers of equal trees among τ_1, \dots, τ_m . This number is closely related to the symmetry coefficient $\sigma(\tau)$, because the product $\kappa(\tau) = \sigma(\tau)\nu(\tau)$ satisfies the recurrence relation

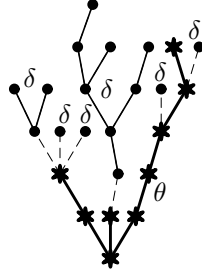
$$\kappa(\tau) = m! \kappa(\tau_1) \cdot \dots \cdot \kappa(\tau_m). \quad (1.32)$$

We introduce the set $OST(\omega)$ of *ordered subtrees* of an ordered tree $\omega \in OT$ by

$$\begin{aligned} OST(\bullet) &= \{\emptyset, \bullet\} \\ OST(\omega) &= \{\emptyset\} \cup \{(\theta_1, \dots, \theta_m) ; \theta_i \in OST(\omega_i)\} \quad \text{for } \omega = (\omega_1, \dots, \omega_m). \end{aligned} \quad (1.33)$$

Each ordered subtree $\theta \in OST(\omega)$ is naturally associated with a tree $\overline{\theta} \in T$ obtained by neglecting the ordering and the \emptyset -components of θ . For every tree $\tau \in T$ we choose, once and for all, an ordering. We denote this ordered tree by $\omega(\tau)$, and we put $OST(\tau) = OST(\omega(\tau))$.

For the tree of Fig. 1.1, considered as an ordered tree, the ordered subtrees correspond to the trees composed of the “star” nodes.



The General Rule. The general composition rule now becomes visible: for $\theta \in OST(\omega)$ we denote by $\omega \setminus \theta$ the “forest” collecting the trees left over when θ has been removed from the ordered tree ω . For brevity we set $\tau \setminus \theta := \omega(\tau) \setminus \theta$. With the conventions $\phi^*(\theta) = \phi^*(\theta)$ and $\phi^*(\emptyset) = 1$ we then have

$$\hat{\phi}(\tau) = \sum_{\theta \in OST(\tau)} \left(\phi^*(\theta) \cdot \prod_{\delta \in \tau \setminus \theta} \phi(\delta) \right). \quad (1.34)$$

This composition formula for the trees up to order 3 reads:

$$\begin{aligned} \hat{\phi}(\bullet) &= \phi^*(\emptyset) \cdot \phi(\bullet) + \phi^*(\bullet) \\ \hat{\phi}(\text{J}) &= \phi^*(\emptyset) \cdot \phi(\text{J}) + \phi^*(\bullet) \cdot \phi(\bullet) + \phi^*(\text{J}) \\ \hat{\phi}(\text{V}) &= \phi^*(\emptyset) \cdot \phi(\text{V}) + \phi^*(\bullet) \cdot \phi(\bullet)^2 + 2\phi^*(\text{J}) \cdot \phi(\bullet) + \phi^*(\text{V}) \\ \hat{\phi}(\text{J}^{\text{J}}) &= \phi^*(\emptyset) \cdot \phi(\text{J}^{\text{J}}) + \phi^*(\bullet) \cdot \phi(\text{J}) + \phi^*(\text{J}) \cdot \phi(\bullet) + \phi^*(\text{J}^{\text{J}}) \end{aligned}$$

The tree $\tau = \text{V}$ has the subtrees displayed in Fig. 1.2. It contains symmetries in that the third and fourth subtrees are topologically equivalent. This explains the factor 2 in the expression for the elementary weight.

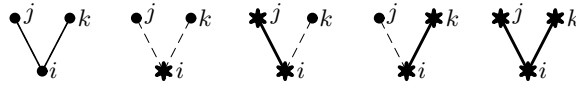


Fig. 1.2. A tree with symmetry

III.1.4 Composition of B-Series

We now extend the above composition law to general B-series, i.e., we insert the B-series themselves into each other, as sketched in Fig. 1.3. This allows us to generalize Lemma 1.9 (because $hf(y)$ is a special B-series).

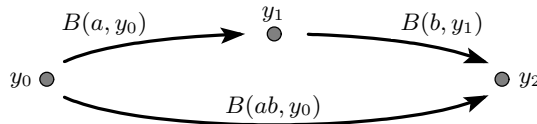


Fig. 1.3. Composition of B-series

We start with an observation of Murua (see, e.g., Murua & Sanz-Serna (1999), p. 1083), namely that the proof of Lemma 1.9 remains the same if the function $hf(y)$ is replaced with any other function $hg(y)$; in this case (1.21) is replaced with

$$hg(B(a, y)) = hg + h^2 a(\bullet) g' f + h^3 a(\bullet) g' f' f + \frac{h^3}{2!} a(\bullet)^2 g''(f, f) + h^4 a(\bullet) a(\bullet) g''(f' f, f) + \dots \quad (1.35)$$

Such series will reappear in Sect. III.3.1 below. Extending this idea further to, say, $f''(y)(v_1, v_2)$, where v_1, v_2 are two fixed vectors, we obtain

$$\begin{aligned} hf''(B(a, y))(v_1, v_2) &= hf''(v_1, v_2) + h^2 a(\bullet) f'''(v_1, v_2, f) \\ &+ h^3 a(\bullet) f'''(v_1, v_2, f' f) + \frac{1}{2!} h^3 a(\bullet)^2 f''''(v_1, v_2, f, f) \\ &+ h^4 a(\bullet) a(\bullet) f''''(v_1, v_2, f' f, f) + \dots \end{aligned} \quad (1.36)$$

This idea will lead to a direct proof of the following theorem of Hairer & Wanner (1974).

Theorem 1.10. *Let $a : T \cup \{\emptyset\} \rightarrow \mathbb{R}$ be a mapping satisfying $a(\emptyset) = 1$ and let $b : T \cup \{\emptyset\} \rightarrow \mathbb{R}$ be arbitrary. Then the B-series $B(a, y)$ inserted into $B(b, \cdot)$ is again a B-series*

$$B(b, B(a, y)) = B(ab, y), \quad (1.37)$$

where the group operation $ab(\tau)$ is as in (1.34), i.e.,

$$ab(\tau) = \sum_{\theta \in OST(\tau)} b(\theta) \cdot a(\tau \setminus \theta) \quad \text{with} \quad a(\tau \setminus \theta) = \prod_{\delta \in \tau \setminus \theta} a(\delta). \quad (1.38)$$

Proof. (a) In part (c) below we prove by induction on $|\vartheta|$, $\vartheta \in T$ that

$$\frac{h^{|\vartheta|}}{\sigma(\vartheta)} F(\vartheta)(B(a, y)) = \sum_{(\tau, \theta) \in A(\vartheta)} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau \setminus \theta) F(\tau)(y), \quad (1.39)$$

where

$$A(\vartheta) = \{(\tau, \theta) ; \tau \in T, \theta \in OST(\tau), \bar{\theta} = \vartheta\}.$$

Multiplying (1.39) by $b(\vartheta)$ and summing over all $\vartheta \in T$ yields the statement (1.37)-(1.38), because

$$\sum_{\vartheta \in T} \sum_{(\tau, \theta) \in A(\vartheta)} \cdot / \cdot = \sum_{\tau \in T} \sum_{\theta \in OST(\tau)} \cdot / \cdot.$$

(b) Choosing a different ordering of τ in the definition of $OST(\tau)$ yields the same sum in (1.39). Therefore (1.39) is equivalent to

$$\frac{h^{|\vartheta|}}{\sigma(\vartheta)} F(\vartheta)(B(a, y)) = \sum_{(\omega, \theta) \in \Omega(\vartheta)} \frac{h^{|\omega|}}{\sigma(\omega)\nu(\omega)} a(\omega \setminus \theta) F(\omega)(y), \quad (1.40)$$

where

$$\Omega(\vartheta) = \{(\omega, \theta) ; \omega \in OT, \theta \in OST(\omega), \bar{\theta} = \vartheta\},$$

and $\nu(\tau)$ is the number of orderings of the tree τ , see (1.31). Functions defined on trees are naturally extended to ordered trees. In (1.40) we use $|\omega| = |\tau|$, $\sigma(\omega) = \sigma(\tau)$, $\nu(\omega) = \nu(\tau)$, $a(\omega \setminus \theta) = a(\tau \setminus \theta)$, and $F(\omega)(y) = F(\tau)(y)$ for $\bar{\omega} = \tau$.

(c) For $\vartheta = \bullet$ and $\omega = (\omega_1, \dots, \omega_m)$ we have $a(\omega \setminus \theta) = a(\omega_1) \cdots a(\omega_m)$ if $\bar{\theta} = \bullet$. Since we have a one-to-one correspondence $(\omega, \theta) \leftrightarrow \omega$ between $\Omega(\bullet)$ and OT , and since the expression in the sum of (1.40) is independent of the ordering of ω , formula (1.40) is precisely Lemma 1.9.

To prove (1.40) for a general tree $\vartheta = [\vartheta_1, \dots, \vartheta_l]$, we apply the idea put forward in (1.36) to $hf^{(l)}(B(a, y))(v_1, \dots, v_l)$ with fixed v_1, \dots, v_l , and obtain as in the proof of Lemma 1.9

$$hf^{(l)}(B(a, y))(v_1, \dots, v_l) = \sum_{m \geq 0} \frac{1}{m!} \sum_{\tau_{l+1} \in T} \cdots \sum_{\tau_{l+m} \in T} \frac{h^{|\tau_{l+1}| + \dots + |\tau_{l+m}| + 1}}{\sigma(\tau_{l+1}) \cdots \sigma(\tau_{l+m})} \\ \cdot a(\tau_{l+1}) \cdots a(\tau_{l+m}) \cdot f^{(l+m)}(y)(v_1, \dots, v_l, F(\tau_{l+1})(y), \dots, F(\tau_{l+m})(y)).$$

Changing the sums over trees to sums over ordered trees we obtain

$$hf^{(l)}(B(a, y))(v_1, \dots, v_l) = \sum_{m \geq 0} \frac{1}{m!} \sum_{\omega_{l+1} \in OT} \cdots \sum_{\omega_{l+m} \in OT} \frac{h^{|\omega_{l+1}| + \dots + |\omega_{l+m}| + 1}}{\kappa(\omega_{l+1}) \cdots \kappa(\omega_{l+m})} \\ \cdot a(\omega_{l+1}) \cdots a(\omega_{l+m}) \cdot f^{(l+m)}(y)(v_1, \dots, v_l, F(\omega_{l+1})(y), \dots, F(\omega_{l+m})(y)).$$

We insert $v_j = \frac{h^{|\vartheta_j|}}{\sigma(\vartheta_j)} F(\vartheta_j)(B(a, y))$ into this relation, and we apply our induction hypothesis

$$v_j = \frac{h^{|\vartheta_j|}}{\sigma(\vartheta_j)} F(\vartheta_j)(B(a, y)) = \sum_{(\omega_j, \theta_j) \in \Omega(\vartheta_j)} \frac{h^{|\omega_j|}}{\kappa(\omega_j)} a(\omega_j \setminus \theta_j) F(\omega_j)(y).$$

We then use the recursive definitions of $\sigma(\vartheta)$ and $F(\vartheta)(y)$ on the left-hand side. On the right-hand side we use the multilinearity of $f^{(l+m)}$, the recursive definitions of $|\omega|$, $\kappa(\omega)$, $F(\omega)(y)$ for $\omega = (\omega_1, \dots, \omega_{l+m})$, and the facts that

$$a(\omega \setminus \theta) = a(\omega_1 \setminus \theta_1) \cdots a(\omega_l \setminus \theta_l) \cdot a(\omega_{l+1}) \cdots a(\omega_{l+m})$$

and

$$\sum_{(\omega_1, \theta_1) \in \Omega(\vartheta_1)} \cdots \sum_{(\omega_l, \theta_l) \in \Omega(\vartheta_l)} \sum_{\omega_{l+1} \in OT} \cdots \sum_{\omega_{l+m} \in OT} \cdot / \cdot = \frac{m! \mu_1! \mu_2! \cdots}{(l+m)!} \sum_{(\omega, \theta) \in \Omega_{l+m}(\vartheta)} \cdot / \cdot$$

where μ_1, μ_2, \dots count equal trees among $\vartheta_1, \dots, \vartheta_l$, and $\Omega_{l+m}(\vartheta)$ consists of those pairs $(\omega, \theta) \in \Omega(\vartheta)$ for which ω is of the form $\omega = (\omega_1, \dots, \omega_{l+m})$. The factorials appear, because to every $(l+m)$ -tuple of the left-hand sum correspond $\binom{l+m}{m, \mu_1, \mu_2, \dots}$ elements in $\Omega_{l+m}(\vartheta)$, obtained by permuting the order. This yields formula (1.40) and hence (1.39). \square

Example 1.11. The composition laws for the trees of order ≤ 4 are

$$\begin{aligned}
ab(\bullet) &= b(\emptyset) \cdot a(\bullet) + b(\bullet) \\
ab(\text{J}) &= b(\emptyset) \cdot a(\text{J}) + b(\bullet) \cdot a(\bullet) + b(\text{J}) \\
ab(\text{V}) &= b(\emptyset) \cdot a(\text{V}) + b(\bullet) \cdot a(\bullet)^2 + 2b(\text{J}) \cdot a(\bullet) + b(\text{V}) \\
ab(\text{J}^{\text{J}}) &= b(\emptyset) \cdot a(\text{J}^{\text{J}}) + b(\bullet) \cdot a(\text{J}) + b(\text{J}) \cdot a(\bullet) + b(\text{J}^{\text{J}}) \\
ab(\text{V}^{\text{V}}) &= b(\emptyset) \cdot a(\text{V}^{\text{V}}) + b(\bullet) \cdot a(\bullet)^3 + 3b(\text{J}) \cdot a(\bullet)^2 + 3b(\text{V}) \cdot a(\bullet) \\
&\quad + b(\text{V}^{\text{V}}) \\
ab(\text{J}^{\text{V}}) &= b(\emptyset) \cdot a(\text{J}^{\text{V}}) + b(\bullet) \cdot a(\bullet)a(\text{J}) + b(\text{J}) \cdot a(\text{J}) + b(\text{J}) \cdot a(\bullet)^2 \\
&\quad + b(\text{V}) \cdot a(\bullet) + b(\text{J}^{\text{J}}) \cdot a(\bullet) + b(\text{J}^{\text{V}}) \\
ab(\text{V}^{\text{J}}) &= b(\emptyset) \cdot a(\text{V}^{\text{J}}) + b(\bullet) \cdot a(\text{V}) + b(\text{J}) \cdot a(\bullet)^2 + 2b(\text{J}^{\text{J}}) \cdot a(\bullet) \\
&\quad + b(\text{V}^{\text{J}}) \\
ab(\text{J}^{\text{J}^{\text{J}}}) &= b(\emptyset) \cdot a(\text{J}^{\text{J}^{\text{J}}}) + b(\bullet) \cdot a(\text{J}^{\text{J}}) + b(\text{J}) \cdot a(\text{J}) + b(\text{J}^{\text{J}}) \cdot a(\bullet) + b(\text{J}^{\text{J}^{\text{J}}})
\end{aligned}$$

Remark 1.12. The composition law (1.38) can alternatively be obtained from the corresponding formula (1.34) for Runge–Kutta methods by using the fact that B-series which represent Runge–Kutta methods are “dense” in the space of all B-series (see Theorem 306A of Butcher 1987).

III.1.5 The Butcher Group



John C. Butcher,
born: 31 March 1933 in Auckland
(New Zealand)

The composition law (1.38) can be turned into a *group operation*, by introducing a *unit element*

$$e(\emptyset) = 1, \quad e(\tau) = 0 \quad \text{for } \tau \in T, \quad (1.41)$$

and by computing the *inverse element* of a given a . This is obtained recursively from the table of Example 1.11, by requiring $aa^{-1}(\tau) = 0$ and by inserting the previously known values of $a^{-1}(\vartheta)$. This gives for the first orders

$$\begin{aligned}
a^{-1}(\bullet) &= -a(\bullet) \\
a^{-1}(\text{J}) &= -a(\text{J}) + a(\bullet)^2 \\
a^{-1}(\text{V}) &= -a(\text{V}) + 2a(\text{J})a(\bullet) - a(\bullet)^3 \\
a^{-1}(\text{J}^{\text{J}}) &= -a(\text{J}^{\text{J}}) + 2a(\text{J})a(\bullet) - a(\bullet)^3
\end{aligned} \tag{1.42}$$

We can distinguish several realizations of this group:

- G_{RK} the set of Runge–Kutta schemes with composition (1.28);
- G_{EW} the set of elementary weights of Runge–Kutta schemes with the composition law (1.34);
- G_{TM} the set of tree mappings $a : T \cup \{\emptyset\} \rightarrow \mathbb{R}$ satisfying $a(\emptyset) = 1$ with composition (1.38);
- G_{BS} the set of B-series (1.23) satisfying $a(\emptyset) = 1$ with composition (1.37).

A technical difficulty concerns the group G_{RK} , where “reducible” schemes must be identified (by deleting unnecessary stages or by combining stages that give identical results) to the same “irreducible” method (see Butcher (1972), or Butcher & Wanner (1996), p. 140). The definition of $\phi(\tau)$ in Theorem 1.4 describes a group isomorphism from G_{RK} to G_{EW} , further, G_{EW} is a subgroup of G_{TM} and Theorem 1.10 shows that formula (1.23) constitutes a group homomorphism from G_{TM} to G_{BS} . Because the elementary differentials are independent (see, e.g., Hairer, Nørsett & Wanner (1993), Exercise 4 of Sect. II.2), the last two groups are isomorphic. The group G_{RK} can also be extended by allowing “continuous” Runge–Kutta schemes with “infinitely many stages” (see Butcher (1972), or Butcher & Wanner (1996), p. 141). The term “Butcher group” was introduced by Hairer & Wanner (1974).

This paper tells the story of a mathematical object that was created by John Butcher in 1972 and was rediscovered by Alain Connes, Henri Moscovici and Dirk Kreimer in 1998. (Ch. Brouder 2004)

Connection with Hopf Algebras and Quantum Field Theory. A surprising connection between Runge–Kutta theory and renormalization in quantum field theory has been discovered by Brouder (2000). One denotes by a *Hopf algebra* a graded algebra which, besides the usual product, also possesses a *coproduct*, a tool used by H. Hopf (1941)² in his topological classification of certain manifolds. Hopf algebras generated by families of rooted trees proved to be extremely useful for simplifying the intricate combinatorics of renormalization (Kreimer 1998). Kreimer’s Hopf algebra \mathcal{H} is the space generated by linear combinations of families of rooted trees and the coproduct is a mapping $\Delta : \mathcal{H} \rightarrow \mathcal{H} \otimes \mathcal{H}$ which is, for the first trees, given by

$$\begin{aligned}
 \Delta(\bullet) &= \bullet \otimes 1 + 1 \otimes \bullet \\
 \Delta(\text{hook}) &= \text{hook} \otimes 1 + \bullet \otimes \bullet + 1 \otimes \text{hook} \\
 \Delta(\text{V}) &= \text{V} \otimes 1 + \bullet \otimes \bullet + 2 \bullet \otimes \text{hook} + 1 \otimes \text{V} \\
 \Delta(\text{hook}^2) &= \text{hook}^2 \otimes 1 + \text{hook} \otimes \bullet + \bullet \otimes \text{hook} + 1 \otimes \text{hook}^2
 \end{aligned} \tag{1.43}$$

It can be clearly seen, that this algebraic structure is precisely the one underlying the composition law of Example 1.11, so that the Butcher group G_{TM} becomes the corresponding *character group*. The so-called *antipodes* of trees $\tau \in \mathcal{H}$, denoted by $S(\tau)$, are for the first trees

² Not to be confused with E. Hopf, the discoverer of the “Hopf bifurcation”.

$$\begin{aligned}
S(\bullet) &= -\bullet \\
S(\text{f}) &= -\text{f} + \bullet\bullet \\
S(\text{V}) &= -\text{V} + 2\text{f}\bullet - \dots \\
S(\text{f}) &= -\text{f} + 2\text{f}\bullet - \dots
\end{aligned} \tag{1.44}$$

and, apparently, describes the *inverse element* (1.42) in the Butcher group.

III.2 Order Conditions for Partitioned Runge–Kutta Methods

We now apply the ideas of the previous section to the creation of the order conditions for partitioned Runge–Kutta methods (II.2.2) of Sect. II.2. These results can then also be applied to Nyström methods.

III.2.1 Bi-Coloured Trees and P-Series

Let us consider a partitioned system

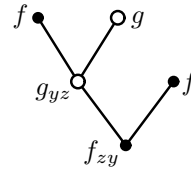
$$\dot{y} = f(y, z), \quad \dot{z} = g(y, z) \tag{2.1}$$

(non-autonomous problems can be brought into this form by appending $\dot{t} = 1$). We start by computing the derivatives of its exact solution, which are to be inserted into the Taylor series expansion. By analogy with (1.4) we obtain in this case the derivatives of y at t_0 as follows:

$$\begin{aligned}
\dot{y} &= f \\
\ddot{y} &= f_y f + f_z g \\
y^{(3)} &= f_{yy}(f, f) + 2f_{yz}(f, g) + f_{zz}(g, g) + f_y f_y f + f_y f_z g + f_z g_y f + f_z g_z g.
\end{aligned} \tag{2.2}$$











Here, f_y, f_z, f_{yz}, \dots denote partial derivatives and all terms are to be evaluated at (y_0, z_0) . Similar expressions are obtained for the derivatives of $z(t)$.

The terms occurring in these expressions are again called the *elementary differentials* $F(\tau)(y, z)$. For their graphical representation as a tree τ , we distinguish between “black” vertices for representing an f and “white” vertices for a g . Upwards pointing branches represent partial derivatives, with respect to y if the branch leads to a black vertex, and with respect to z if it leads to a white vertex. With this convention, the graph to the right corresponds to the expression $f_{zy}(g_{yz}(f, g), f)$ (see Table 2.1 for more examples).



We denote by TP the set of graphs obtained by the above procedure, and we call them (rooted) *bi-coloured trees*. The first graphs are \bullet and \circ . By analogy with Definition 1.1, we denote by

Table 2.1. Bi-coloured trees, elementary differentials, and coefficients

$ \tau $	τ	graph	$\alpha(\tau)$	$F(\tau)$	$\gamma(\tau)$	$\phi(\tau)$	$\sigma(\tau)$
1	\bullet	\bullet	1	f	1	$\sum_i b_i$	1
2	$[\bullet]_y$		1	$f_y f$	2	$\sum_{ij} b_i a_{ij}$	1
2	$[\circ]_y$		1	$f_z g$	2	$\sum_{ij} b_i \hat{a}_{ij}$	1
3	$[\bullet, \bullet]_y$		1	$f_{yy}(f, f)$	3	$\sum_{ijk} b_i a_{ij} a_{ik}$	2
3	$[\bullet, \circ]_y$		2	$f_{yz}(f, g)$	3	$\sum_{ijk} b_i a_{ij} \hat{a}_{ik}$	1
3	$[\circ, \circ]_y$		1	$f_{zz}(g, g)$	3	$\sum_{ijk} b_i \hat{a}_{ij} \hat{a}_{ik}$	2
3	$[[\bullet]_y]_y$		1	$f_y f_y f$	6	$\sum_{ijk} b_i a_{ij} a_{jk}$	1
3	$[[\circ]_y]_y$		1	$f_y f_z g$	6	$\sum_{ijk} b_i a_{ij} \hat{a}_{jk}$	1
3	$[[\bullet]_z]_y$		1	$f_z g_y f$	6	$\sum_{ijk} b_i \hat{a}_{ij} a_{jk}$	1
3	$[[\circ]_z]_y$		1	$f_z g_z g$	6	$\sum_{ijk} b_i \hat{a}_{ij} \hat{a}_{jk}$	1
1	\circ	\circ	1	g	1	$\sum_i \hat{b}_i$	1
2	$[\bullet]_z$		1	$g_y f$	2	$\sum_{ij} \hat{b}_i a_{ij}$	1
	etc	etc		etc		etc	

$$[\tau_1, \dots, \tau_m]_y \quad \text{and} \quad [\tau_1, \dots, \tau_m]_z, \quad \tau_1, \dots, \tau_m \in TP$$

the bi-coloured trees obtained by connecting the roots of τ_1, \dots, τ_m to a new root, which is \bullet in the first case, and \circ in the second. Furthermore, we denote by TP_y and TP_z the subsets of TP which are formed by trees with black and white roots, respectively. Hence, the trees of TP_y correspond to derivatives of $y(t)$, whereas those of TP_z correspond to derivatives of $z(t)$.

As in Definition 1.2 we denote the number of vertices of $\tau \in TP$ by $|\tau|$, the order of τ . The symmetry coefficient $\sigma(\tau)$ is again defined by

$$\sigma(\bullet) = \sigma(\circ) = 1,$$

and, for $\tau = [\tau_1, \dots, \tau_m]_y$ or $\tau = [\tau_1, \dots, \tau_m]_z$, by

$$\sigma(\tau) = \sigma(\tau_1) \cdot \dots \cdot \sigma(\tau_m) \cdot \mu_1! \mu_2! \dots, \quad (2.3)$$

where the integers μ_1, μ_2, \dots count equal trees among $\tau_1, \dots, \tau_m \in TP$. This is formally the same definition as in Sect. III.1. Observe, however, that $\sigma(\tau)$ depends on the colouring of the vertices. For example, we have $\sigma(\mathbf{V}) = 2$, but $\sigma(\mathbf{V}^\circ) = 1$. By analogy with Definition 1.8 we have:

Definition 2.1 (P-Series). For a mapping $a : TP \cup \{\emptyset_y, \emptyset_z\} \rightarrow \mathbb{R}$ a series of the form

$$P(a, (y, z)) = \begin{pmatrix} a(\emptyset_y)y + \sum_{\tau \in TP_y} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau) F(\tau)(y, z) \\ a(\emptyset_z)z + \sum_{\tau \in TP_z} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau) F(\tau)(y, z) \end{pmatrix}$$

is called a *P-series*.

The following results correspond to Lemma 1.9 and formula (1.26). They are obtained in exactly the same manner as the corresponding results for non-partitioned Runge–Kutta methods (Sect. III.1). We therefore omit their proofs.

Lemma 2.2. Let $a : TP \cup \{\emptyset_y, \emptyset_z\} \rightarrow \mathbb{R}$ satisfy $a(\emptyset_y) = a(\emptyset_z) = 1$. Then

$$h \begin{pmatrix} f(P(a, (y, z))) \\ g(P(a, (y, z))) \end{pmatrix} = P(a', (y, z)),$$

where $a'(\emptyset_y) = a'(\emptyset_z) = 0$, $a'(\bullet) = a'(\circ) = 1$, and

$$a'(\tau) = a(\tau_1) \cdot \dots \cdot a(\tau_m), \quad (2.4)$$

if either $\tau = [\tau_1, \dots, \tau_m]_y$ or $\tau = [\tau_1, \dots, \tau_m]_z$. \square

Theorem 2.3 (P-Series of Exact Solution). The exact solution of (2.1) is a P-series $(y(t_0 + h), z(t_0 + h)) = P(\mathbf{e}, (y_0, z_0))$, where $\mathbf{e}(\emptyset_y) = \mathbf{e}(\emptyset_z) = 1$ and

$$\mathbf{e}(\tau) = \frac{1}{\gamma(\tau)} \quad \text{for all } t \in TP \quad (2.5)$$

where the $\gamma(\tau)$ have the same values as for mono-coloured trees. \square

III.2.2 Order Conditions for Partitioned Runge–Kutta Methods

The next result corresponds to Theorem 1.4 and is a consequence of Lemma 2.2.

Theorem 2.4 (P-Series of Numerical Solution). The numerical solution of a partitioned Runge–Kutta method (II.2.2) is a P-series $(y_1, z_1) = P(\phi, (y_0, z_0))$, where $\phi(\emptyset_y) = \phi(\emptyset_z) = 1$ and

$$\phi(\tau) = \begin{cases} \sum_{i=1}^s b_i \phi_i(\tau) & \text{for } \tau \in TP_y \\ \sum_{i=1}^s \hat{b}_i \phi_i(\tau) & \text{for } \tau \in TP_z. \end{cases} \quad (2.6)$$

The expression $\phi_i(\tau)$ is defined by $\phi_i(\bullet) = \phi_i(\circ) = 1$ and by

$$\phi_i(\tau) = \psi_i(\tau_1) \dots \psi_i(\tau_m) \quad \text{with} \quad \psi_i(\tau_k) = \begin{cases} \sum_{j_k=1}^s a_{ij_k} \phi_{j_k}(\tau_k) & \text{if } \tau_k \in TP_y \\ \sum_{j_k=1}^s \hat{a}_{ij_k} \phi_{j_k}(\tau_k) & \text{if } \tau_k \in TP_z \end{cases} \quad (2.7)$$

for $\tau = [\tau_1, \dots, \tau_m]_y$ or $\tau = [\tau_1, \dots, \tau_m]_z$.

Proof. These formulas result from Lemma 2.2 by writing $(hk_i, h\ell_i)$ from the formulas (II.2.2) as a P-series $(hk_i, h\ell_i) = P(\phi_i, (y_0, z_0))$ so that

$$(h \sum_j a_{ij} k_j, h \sum_j \hat{a}_{ij} \ell_j) = P(\psi_i, (y_0, z_0))$$

is also a P-series. Observe that equation (2.6) corresponds to (1.16) (where \mathbf{g}_i has to be replaced with ϕ_i) and that formula (2.7) comprises (1.13) and (1.15), where we now write ψ_i instead of \mathbf{u}_i . \square

The expressions $\phi(\tau)$ are shown in Table 2.1 for all trees in TP_y up to order $|\tau| \leq 3$. A similar table must be added for trees in TP_z , where all roots are white and all b_i are replaced with \hat{b}_i . The general rule is the following: attach to every vertex a summation index. Then, the expression $\phi(\tau)$ is a sum over all summation indices with the summand being a product of b_i or \hat{b}_i (depending on whether the root “ i ” is black or white) and of a_{jk} (if “ k ” is black) or \hat{a}_{jk} (if “ k ” is white), for each vertex “ k ” directly above “ j ”.

Theorem 2.5 (Order Conditions). *A partitioned Runge–Kutta method (II.2.2) has order r , i.e., $y_1 - y(t_0 + h) = \mathcal{O}(h^{r+1})$, $z_1 - z(t_0 + h) = \mathcal{O}(h^{r+1})$, if and only if*

$$\phi(\tau) = \frac{1}{\gamma(\tau)} \quad \text{for } \tau \in TP_y \cup TP_z \text{ with } |\tau| \leq r. \quad (2.8)$$

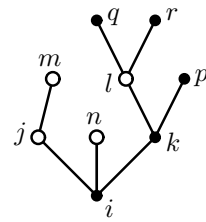
Proof. This corresponds to Theorem 1.5 and is seen by comparing the expansions of Theorems 2.4 and 2.3. \square

Example 2.6. We see that not only does every individual Runge–Kutta method have to be of order r , but also the so-called *coupling conditions* between the coefficients of both methods must hold. The order conditions mentioned above (see formulas (II.2.3) and (II.2.5)) correspond to the trees \mathcal{J} , \mathcal{J} , \mathcal{J} and \mathcal{J} . For the tree sketched below we obtain

$$\sum_{i,j,k,l,m,n,p,q,r} b_i \hat{a}_{ij} \hat{a}_{jm} \hat{a}_{in} a_{ik} \hat{a}_{kl} a_{lq} a_{lr} a_{kp} = \frac{1}{9 \cdot 2 \cdot 5 \cdot 3}$$

or, by using $\sum_j a_{ij} = c_i$ and $\sum_j \hat{a}_{ij} = \hat{c}_i$,

$$\sum_{i,j,k,l} b_i \hat{c}_i \hat{a}_{ij} \hat{c}_j a_{ik} c_k \hat{a}_{kl} c_l^2 = \frac{1}{270}.$$



III.2.3 Order Conditions for Nyström Methods

A “modern” order theory for Nyström methods (II.2.11) of Sect. II.2.3 was first given in 1976 by Hairer & Wanner (see Sect. II.14 of Hairer, Nørsett & Wanner

1993). Later it turned out that these conditions are obtained easily by applying the theory of partitioned Runge–Kutta methods to the system

$$\dot{y} = z \quad \dot{z} = g(y, z), \quad (2.9)$$

which is of the form (2.1). This function has the partial derivative $f_z = I$ and all other derivatives of f are zero. As a consequence, many elementary differentials are zero and the corresponding order conditions can be omitted. The only trees remaining are those for which

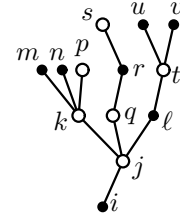
$$\text{“black vertices have at most one son and this son must be white”}. \quad (2.10)$$

Example 2.7. The tree sketched below apparently satisfies condition (2.10) and the corresponding order condition becomes, by Theorem 2.4 and formula (2.8),

$$\sum_{i,j,k,\dots,v} b_i \hat{a}_{ij} \hat{a}_{jk} a_{km} a_{kn} \hat{a}_{kp} \hat{a}_{jq} a_{qr} \hat{a}_{rs} a_{jt} \hat{a}_{\ell t} a_{tu} a_{tv} = \frac{1}{13 \cdot 12 \cdot 4 \cdot 3 \cdot 2 \cdot 4 \cdot 3}.$$

Due to property (2.10), each a_{ik} inside the tree comes with a corresponding \hat{a}_{kj} , and by (2.10), both factors contract to an \bar{a}_{ij} ; similarly, the black root is only connected to one white vertex, the corresponding $b_i \hat{a}_{ij}$ simplifies to \bar{b}_j . We thus get

$$\sum_{j,k,q,s,t} \bar{b}_j \hat{a}_{jk} c_k^2 \hat{a}_{jq} \bar{a}_{qs} \bar{a}_{jt} c_t^2 = \frac{1}{13 \cdot 3456}.$$



Each of the above order conditions for a tree in TP_y has a “twin” in TP_z of one order lower with the root cut off. For the above example this twin becomes

$$\sum_{j,k,q,s,t} b_j \hat{a}_{jk} c_k^2 \hat{a}_{jq} \bar{a}_{qs} \bar{a}_{jt} c_t^2 = \frac{1}{3456}.$$

We need only consider the trees in TP_z if

$$\bar{b}_i = b_i(1 - c_i)$$

is satisfied (see Lemma II.14.13 of Hairer, Nørsett & Wanner (1993), Sect. II.14).

Remark 2.8. Strictly speaking, the theory of partitioned methods is applicable to Nyström methods only if the matrix (\hat{a}_{ij}) is invertible. However, since we arrive at expansions with a finite number of algebraic conditions, we can recover the singular case by a continuous perturbation of the coefficients.

Equations without Friction. Although condition (2.10) already eliminates many order conditions, Nyström methods for the general problem $\ddot{y} = g(y, \dot{y})$ cannot be much better than an excellent Runge–Kutta method applied pairwise to system (2.9).

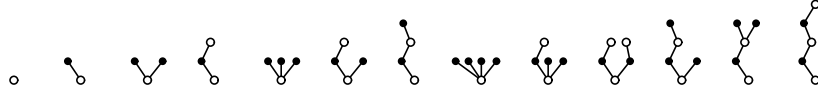
There *is*, however, an important special case where much more progress is possible, namely equations of the type

$$\ddot{y} = g(y), \quad (2.11)$$

which corresponds to motion without friction. In this case, the function for \dot{z} in (2.9) is *independent of z* , and in addition to (2.10) we have a second condition, namely

$$\text{“white vertices have only black sons”}. \quad (2.12)$$

Both conditions reduce the remaining trees drastically. Along each branch, there occur alternating black and white vertices. Ramifications only happen at white vertices. This case allows the construction of excellent numerical methods of high orders. For example, the following 13 trees



assure order 5, whereas ordinary Runge–Kutta theory requires 17 conditions for this order. See Hairer, Nørsett & Wanner (1993), pages 291f, for tables, examples and references.

III.3 Order Conditions for Composition Methods

We have seen in the preceding chapter that composition methods of arbitrarily high order can be obtained with the use of Theorem II.4.1. However, as demonstrated in Fig. II.4.4, these methods are not attractive for high orders. This section is devoted to the derivation of order conditions, which then allow the construction of optimal high order composition methods.

The order conditions for these methods are often derived via the Baker–Campbell–Hausdorff formula. This will be the subject of Sect. III.5 below. Only very recently, Murua & Sanz-Serna (1999) have found an elegant theory based on the idea of B-series. This paper has largely inspired the subsequent presentation.

III.3.1 Introduction

The principal tool in this section is the Taylor series expansion

$$\Phi_h(y) = y + h d_1(y) + h^2 d_2(y) + h^3 d_3(y) + \dots \quad (3.1)$$

of the basic method. The only hypothesis which we require for this method is *consistency*, i.e., that

$$d_1(y) = f(y). \quad (3.2)$$

All other functions $d_i(y)$ are arbitrary.

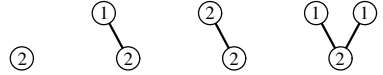
The underlying idea for obtaining the expansions for composition methods is, in fact, very simple: we just insert the series (3.1), with varying values of h , into itself. All our experience from Sect. III.1.2 with the insertion of a B-series into a function will certainly be helpful. We demonstrate this for the case of the composition $\Psi_h = \Phi_{\alpha_2 h} \circ \Phi_{\alpha_1 h}$. Applied to an initial value y_0 , this gives with (3.1)

$$\begin{aligned} y_1 &= \Phi_{\alpha_1 h}(y_0) = y_0 + h\alpha_1 d_1(y_0) + h^2\alpha_1^2 d_2(y_0) + \dots \\ y_2 &= \Phi_{\alpha_2 h}(y_1) = y_1 + h\alpha_2 d_1(y_1) + h^2\alpha_2^2 d_2(y_1) + \dots \end{aligned} \quad (3.3)$$

We now insert the first series into the second, in the same way as we did in (1.35). Then, for example, the term $h^2\alpha_2^2 d_2(y_1)$ becomes

$$\begin{aligned} y_2 = \dots &+ h^2\alpha_2^2 d_2(y_0) + h^3\alpha_2^2\alpha_1 d_2'(y_0)d_1(y_0) \\ &+ h^4\alpha_2^2\alpha_1^2 d_2'(y_0)d_2(y_0) + \frac{h^4}{2}\alpha_2^2\alpha_1^2 d_2''(y_0)(d_1(y_0), d_1(y_0)) + \dots \end{aligned} \quad (3.4)$$

We see that we arrive at “generalized” B-series, where the elementary differentials contain not only *one* function, but are composed of *infinitely many* functions and their derivatives. We symbolize the four terms written in (3.4) by the trees



This leads us to the following definition.

Definition 3.1 (∞ -Trees, B_∞ -series). We extend Definitions 1.1 and 1.2 to T_∞ , the set of all rooted trees where each vertex bears a positive integer without any further restriction, and use the notation

- ①, ②, ③, ... = the trees with one vertex;
- $[\tau_1, \dots, \tau_m]_i$ = the tree τ formed by a new root ① connected to τ_1, \dots, τ_m ;
- $F(\textcircled{i})(y) = d_i(y)$;
- $F(\tau)(y) = d_i^{(m)}(y)(F(\tau_1)(y), \dots, F(\tau_m)(y))$ for τ as above;
- $|\tau| = 1 + |\tau_1| + \dots + |\tau_m|$, the number of vertices of τ ;
- $||\tau|| = i + ||\tau_1|| + \dots + ||\tau_m||$, the sum of the labels of τ ;
- $\sigma(\tau) = \mu_1! \mu_2! \cdot \dots \cdot \sigma(\tau_1) \cdot \dots \cdot \sigma(\tau_m)$,
where μ_1, μ_2, \dots count equal trees among τ_1, \dots, τ_m ,
the symmetry coefficient respecting the labels;
- $i(\tau) = i$, the label of the root of τ .

For a map $a : T_\infty \cup \{\emptyset\} \rightarrow \mathbb{R}$ we write

$$B_\infty(a, y) = a(\emptyset)y + \sum_{\tau \in T_\infty} \frac{h^{||\tau||}}{\sigma(\tau)} a(\tau) F(\tau)(y) \quad (3.5)$$

which extends the notion of B-series to the new situation.

Example 3.2. For the tree

$$\tau = \begin{array}{c} \textcircled{5} \textcircled{6} \textcircled{6} \\ | \quad | \quad | \\ \textcircled{1} \textcircled{7} \\ | \quad | \\ \textcircled{4} \end{array} \Leftrightarrow \tau = [\tau_1, \tau_2]_4 \quad \text{where} \quad \tau_1 = \textcircled{1}, \quad \tau_2 = \begin{array}{c} \textcircled{5} \textcircled{6} \textcircled{6} \\ | \quad | \quad | \\ \textcircled{7} \end{array} \quad (3.6)$$

we have

$$F(\tau)(y) = d_4''(y) \left(d_1(y), d_7'''(y) (d_5(y), d_6(y), d_6(y)) \right)$$

$$\tau = [\textcircled{1}, [\textcircled{5}, \textcircled{6}, \textcircled{6}]_7]_4, \quad |\tau| = 6, \quad ||\tau|| = 29, \quad \sigma(\tau) = 2, \quad i(\tau) = 4.$$

The above calculations for (3.4) are governed by the following lemma.

Lemma 3.3. For a series $B_\infty(a, y)$ with $a(\emptyset) = 1$ we have

$$h^i d_i \left(B_\infty(a, y) \right) = \sum_{\tau \in T_\infty, i(\tau)=i} \frac{h^{||\tau||}}{\sigma(\tau)} a'(\tau) F(\tau)(y), \quad (3.7)$$

where $a'(\textcircled{i}) = 1$ and

$$a'(\tau) = a(\tau_1) \cdot \dots \cdot a(\tau_m) \quad \text{for } \tau = [\tau_1, \dots, \tau_m]_i. \quad (3.8)$$

Proof. This is a straightforward extension of Lemma 1.9 with exactly the same proof. \square

The preceding lemma leads directly to the order conditions for composition methods. However, if we continue with compositions of the type (II.4.1), we arrive at conditions without real solutions. We therefore turn to compositions including the adjoint method as well.

III.3.2 The General Case

As in (II.4.6), we consider

$$\Psi_h = \Phi_{\alpha_s h} \circ \Phi_{\beta_s h}^* \circ \dots \circ \Phi_{\alpha_2 h} \circ \Phi_{\beta_2 h}^* \circ \Phi_{\alpha_1 h} \circ \Phi_{\beta_1 h}^*, \quad (3.9)$$

and we obtain with the help of the above lemma the corresponding B_∞ -series.

Lemma 3.4 (Recurrence Relations). The following compositions are B_∞ -series

$$\begin{aligned} (\Phi_{\beta_k h}^* \circ \dots \circ \Phi_{\alpha_1 h} \circ \Phi_{\beta_1 h}^*)(y) &= B_\infty(b_k, y) \\ (\Phi_{\alpha_k h} \circ \Phi_{\beta_k h}^* \circ \dots \circ \Phi_{\alpha_1 h} \circ \Phi_{\beta_1 h}^*)(y) &= B_\infty(a_k, y). \end{aligned} \quad (3.10)$$

Their coefficients are recursively given by $a_k(\emptyset) = 1$, $b_k(\emptyset) = 1$, $a_0(\tau) = 0$ for all $\tau \in T_\infty$, and

$$\begin{aligned} b_k(\tau) &= a_{k-1}(\tau) - (-\beta_k)^{i(\tau)} b'_k(\tau), \\ a_k(\tau) &= b_k(\tau) + \alpha_k^{i(\tau)} b'_k(\tau). \end{aligned} \quad (3.11)$$

Proof. The coefficients $a_0(\tau)$ correspond to the identity map $B_\infty(a_0, y) = y$. The second formula of (3.11) follows from

$$B_\infty(a_k, y) = \Phi_{\alpha_k h} \left(B_\infty(b_k, y) \right) = B_\infty(b_k, y) + \sum_{i \geq 1} \alpha_k^i h^i d_i \left(B_\infty(b_k, y) \right),$$

and from an application of Lemma 3.3.

The relation $B_\infty(b_k, y) = \Phi_{\beta_k h}^* (B_\infty(a_{k-1}, y))$, which involves the adjoint method, needs a little trick: we write it as $B_\infty(a_{k-1}, y) = \Phi_{-\beta_k h} (B_\infty(b_k, y))$ (remember that $\Phi_h^* = \Phi_{-h}^{-1}$), apply Lemma 3.3 again, and reverse the formula. This gives the first equation of (3.11). \square

Adding the equations of (3.11), we get

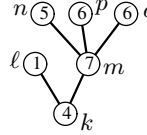
$$a_k(\tau) = a_{k-1}(\tau) + (\alpha_k^{i(\tau)} - (-\beta_k)^{i(\tau)}) b'_k(\tau). \quad (3.12)$$

Because of $b'_k(\textcircled{i}) = 1$, we obtain

$$\begin{aligned} a_k(\textcircled{i}) &= \sum_{\ell=1}^k (\alpha_\ell^i - (-\beta_\ell)^i) \\ b_k(\textcircled{i}) &= \sum_{\ell=1}^{k-1} \alpha_\ell^i - \sum_{\ell=1}^k (-\beta_\ell)^i = \sum_{\ell=1}^k{}' (\alpha_\ell^i - (-\beta_\ell)^i). \end{aligned} \quad (3.13)$$

The fact that, for $b_k(\textcircled{i})$, the sum of $(-\beta_\ell)^i$ is from 1 to k , but the sum of α_ℓ^i is only from 1 to $k-1$, has been *indicated by a prime* attached to the summation symbol. Continuing to apply the formulas (3.11) and (3.12) to more and more complicated trees, we quickly understand the general rule for the coefficients of an arbitrary tree.

Example 3.5. The tree τ in (3.6) gives



$$a_s(\tau) = \sum_{k=1}^s (\alpha_k^4 - \beta_k^4) \sum_{\ell=1}^k{}' (\alpha_\ell + \beta_\ell) \cdot \sum_{m=1}^k{}' (\alpha_m^7 + \beta_m^7) \sum_{n=1}^m{}' (\alpha_n^5 + \beta_n^5) \left(\sum_{p=1}^m{}' (\alpha_p^6 - \beta_p^6) \right)^2. \quad (3.14)$$

The Order Conditions. The exact solution of $\dot{y} = f(y)$ is a B -series $y(t_0 + h) = B(\mathbf{e}, y_0)$ (see (1.26)). Since $d_1(y) = f(y)$, every B -series is also a B_∞ -series with $\mathbf{e}(\tau) = 0$ for trees with at least one label different from 1. Therefore, we also have $y(t_0 + h) = B_\infty(\mathbf{e}, y_0)$, where the coefficients $\mathbf{e}(\tau)$ satisfy $\mathbf{e}(\textcircled{1}) = 1$, $\mathbf{e}(\tau) = 0$ if $i(\tau) > 1$, and

$$\mathbf{e}(\tau) = \frac{1}{|\tau|} \mathbf{e}(\tau_1) \cdot \dots \cdot \mathbf{e}(\tau_m) \quad \text{for } \tau = [\tau_1, \dots, \tau_m]_1. \quad (3.15)$$

Theorem 3.6. *The composition method $\Psi_h(y) = B_\infty(a_s, y)$ of (3.9) has order p if*

$$a_s(\tau) = \mathbf{e}(\tau) \quad \text{for } \tau \in T_\infty \text{ with } \|\tau\| \leq p. \quad (3.16)$$

Proof. This follows from a comparison of the B_∞ -series for the numerical and the exact solution. For the necessity of (3.16), the independence of the elementary differentials has to be studied as in Exercise 3. \square

III.3.3 Reduction of the Order Conditions

The order conditions of the foregoing section are indeed beautiful, but for the moment they are not of much use, because of the enormous number of trees in T_∞ of a certain order. For example, there are 166 trees in T_∞ with $\|\tau\| \leq 6$. Fortunately, the equations are not all independent, as we shall see now.

Definition 3.7 (Butcher 1972, Murua & Sanz-Serna 1999). For two trees in T_∞ , $u = [u_1, \dots, u_m]_i$ and $v = [v_1, \dots, v_l]_j$, we denote

$$u \circ v := [u_1, \dots, u_m, v]_i, \quad u \times v := [u_1, \dots, u_m, v_1, \dots, v_l]_{i+j} \quad (3.17)$$

and call them the *Butcher product* and *merging product*, respectively (see Fig. 3.1).

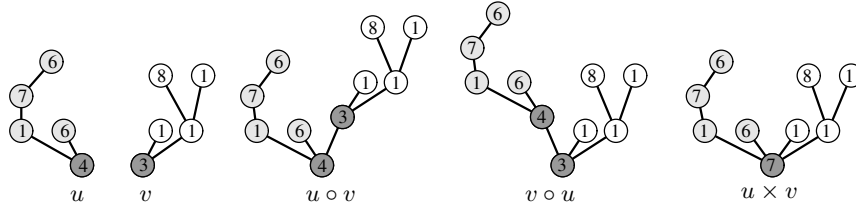


Fig. 3.1. The Butcher product and the merging product

The merging product is associative and commutative, the Butcher product is neither of the two. To simplify the notation, we write products of *several* factors without parentheses, when we mean evaluation from left to right:

$$u \circ v_1 \circ v_2 \circ \dots \circ v_s = (((u \circ v_1) \circ v_2) \circ \dots) \circ v_s. \quad (3.18)$$

Here the factors v_1, \dots, v_s can be freely permuted.

All subsequent results concern properties of $a_k(\tau)$ as well as $b_k(\tau)$, valid for all k . To avoid writing all formulas twice, we replace $a_k(\tau)$ and $b_k(\tau)$ everywhere by a neutral symbol $c(\tau)$.

Lemma 3.8 (Switching Lemma). *All a_k, b_k of Lemma 3.4 satisfy, for all $u, v \in T_\infty$, the relation*

$$c(u \circ v) + c(v \circ u) = c(u) \cdot c(v) - c(u \times v). \quad (3.19)$$

Proof. The recursion formulas (3.11) are of the form

$$a(\tau) = b(\tau) + \alpha^{i(\tau)} b'(\tau). \quad (3.20)$$

We arrange this formula, for all five trees of Fig. 3.1, as follows:

$$\begin{aligned} & a(u \circ v) + a(v \circ u) + a(u \times v) - a(u)a(v) \\ = & b(u \circ v) + b(v \circ u) + b(u \times v) - b(u)b(v) \\ & + \alpha^{i(u)} b'(u \circ v) + \alpha^{i(v)} b'(v \circ u) + \alpha^{i(u)+i(v)} b'(u \times v) \\ & - \alpha^{i(u)} b'(u)b(v) - \alpha^{i(v)} b'(v)b(u) - \alpha^{i(u)} \alpha^{i(v)} b'(u)b'(v). \end{aligned}$$

Because of $b'(u \circ v) = b'(u)b(v)$ and $b'(u \times v) = b'(u)b'(v)$, the last two rows cancel, hence

$$a(\tau) \text{ satisfies (3.19)} \Leftrightarrow b(\tau) \text{ satisfies (3.19)}. \quad (3.21)$$

Thus, beginning with a_0 , then b_1 , then a_1 , etc., all a_k and b_k must satisfy (3.19). \square

The Switching Lemma 3.8 reduces considerably the number of order conditions. Since the right-hand expression involves only trees with $|\tau| < |u \circ v|$, and since relation (3.19) is also satisfied by $e(\tau)$, an induction argument shows that the order conditions (3.16) for the trees $u \circ v$ and $v \circ u$ are equivalent. The operation $u \circ v \mapsto v \circ u$ consists simply in switching the root from one vertex to the next. By repeating this argument, we see that we can freely move the root inside the graph, and of all these trees, only one needs to be retained. For order 6, for example, there remain 68 conditions out of the original 166.

Our next results show how relation (3.19) also generates a considerable amount of reductions of the order conditions. These ideas (for the special situation of symplectic methods) have already been exploited by Calvo & Hairer (1995b).

Lemma 3.9. *Assume that all b_k of Lemma 3.4 satisfy a relation of the form*

$$\sum_{i=1}^N A_i \prod_{j=1}^{m_i} c(u_{ij}) = 0 \quad (3.22)$$

with all $m_i > 0$. Then, for any tree w , all a_k and b_k satisfy the relation

$$\sum_{i=1}^N A_i c(w \circ u_{i1} \circ u_{i2} \circ \dots \circ u_{i,m_i}) = 0. \quad (3.23)$$

Proof. The relation (3.20), written for the tree $w \circ u_{i1} \circ u_{i2} \circ \dots \circ u_{i,m_i}$, is

$$\begin{aligned} a(w \circ u_{i1} \circ \dots \circ u_{i,m_i}) &= b(w \circ u_{i1} \circ \dots \circ u_{i,m_i}) \\ &+ \alpha^{i(w)} b'(w) b(u_{i1}) \cdot \dots \cdot b(u_{i,m_i}). \end{aligned}$$

Multiplying with A_i and summing over i , this shows that, under the hypothesis (3.22) for b , the relation (3.23) holds for b if and only if it holds for a . The coefficients $a_0(\tau) = 0$ for the identity map satisfy (3.22) and (3.23) because $m_i > 0$. Starting from this, we again conclude (3.23) recursively for all a_k and b_k . \square

The following lemma³ extends formula (3.19) to the case of *several* factors.

Lemma 3.10. *For any three trees u, v, w all a_k, b_k of Lemma 3.4 satisfy a relation*

$$c(u \circ v \circ w) + c(v \circ u \circ w) + c(w \circ u \circ v) = c(u) \cdot c(v) \cdot c(w) + \dots, \quad (3.24)$$

where the dots indicate a linear combination of products $\prod_j c(v_j)$ with $|v_1| + |v_2| + \dots < |u| + |v| + |w|$ and, for each term, at least one of the v_j possesses a label larger than one. The general formula, for m trees u_1, \dots, u_m , is

$$\sum_{i=1}^m c(u_i \circ u_1 \circ \dots \circ u_{i-1} \circ u_{i+1} \circ \dots \circ u_m) = \prod_{i=1}^m c(u_i) + \dots \quad (3.25)$$

Proof. We apply Lemma 3.9 to (3.19) and obtain

$$c(w \circ (u \circ v)) + c(w \circ (v \circ u)) = c(w \circ u \circ v) - c(w \circ (u \times v)). \quad (3.26)$$

Next, we apply the Switching Lemma 3.8 to the trees to the left and get

$$\begin{aligned} c(w \circ (u \circ v)) + c(u \circ v \circ w) &= c(w) \cdot c(u \circ v) - c(w \times (u \circ v)) \\ c(w \circ (v \circ u)) + c(v \circ u \circ w) &= c(w) \cdot c(v \circ u) - c(w \times (v \circ u)). \end{aligned}$$

Adding these formulas and subtracting (3.26) gives

$$c(u \circ v \circ w) + c(v \circ u \circ w) + c(w \circ u \circ v) = c(w)(c(u \circ v) + c(v \circ u)) + \dots$$

which becomes (3.24) after another use of the Switching Lemma. Thereby, everything which goes into “+ ...” contains somewhere a merging product, whose roots introduce necessarily labels larger than one.

Continuing like this, we get recursively (3.25) for all m . \square

In order that the further simplifications do not turn into chaos, we fix, once and for all, a *total order relation* (written $<$) on T_∞ , where we only require that the order respects the number of vertices, i.e., that

$$u < v \quad \text{whenever} \quad |u| < |v|. \quad (3.27)$$

Similar to the strategy introduced by Hall (1950) for simplifying bracket expressions in Lie algebras, we define the following subset of T_∞ .

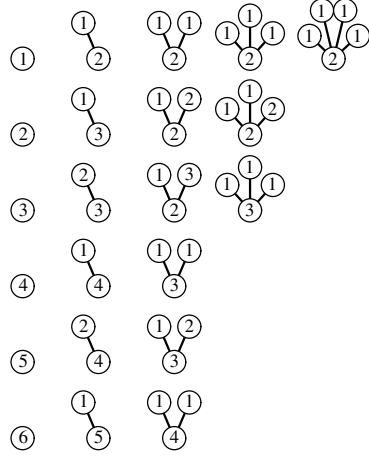
³ due to A. Murua, private communication, Feb. 2001

Definition 3.11 (Hall Set). The *Hall set* corresponding to an order relation (3.27) is a subset $\mathcal{H} \subset T_\infty$ defined by

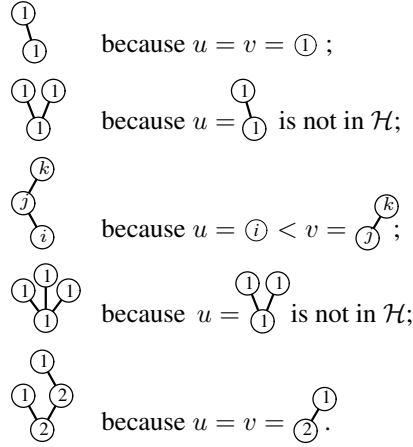
$$\begin{aligned} \textcircled{i} &\in \mathcal{H} \quad \text{for } i = 1, 2, 3, \dots \\ \tau \in \mathcal{H} &\Leftrightarrow \text{there exist } u, v \in \mathcal{H}, u > v, \text{ such that } \tau = u \circ v. \end{aligned}$$

Example 3.12. The trees in the subsequent table are ordered from left to right with respect to $|\tau|$, and from top to bottom within fixed $|\tau|$. There remain finally 22 conditions for order 6.

A Hall set \mathcal{H} with $||\tau|| \leq 6$:



Not in \mathcal{H} are, for example:



Theorem 3.13 (Murua & Sanz-Serna 1999). For each $\tau \in T_\infty$ there are constants A_i , integers m_i and trees $u_{ij} \in \mathcal{H}$ such that for all a_k, b_k of Lemma 3.4 we have

$$c(\tau) = \sum_{i=1}^N A_i \prod_{j=1}^{m_i} c(u_{ij}), \quad u_{ij} \in \mathcal{H}, \quad |u_{i1}| + \dots + |u_{i,m_i}| \leq |\tau|. \quad (3.28)$$

Proof. We proceed by induction on $|\tau|$. For $\tau = \textcircled{i}$ the statement is trivial, because $\textcircled{i} \in \mathcal{H}$. We thus consider $\tau \in T_\infty$ with $|\tau| \geq 2$, write it as $\tau = u \circ v$, and conclude through the following two steps.

First Step. We apply the induction hypothesis (3.28) to v , i.e.,

$$c(v) = \sum_i B_i \prod_j c(v_{ij}), \quad v_{ij} \in \mathcal{H}, \quad \sum_j |v_{ij}| \leq |v|. \quad (3.29)$$

To this, we apply Lemma 3.9 followed by the Switching Lemma 3.8:

$$\begin{aligned} c(\tau) &= c(u \circ v) = \sum_i B_i c(u \circ v_{i1} \circ v_{i2} \dots \circ v_{i,n_i}) \\ &= - \sum_i B_i c(v_{in_i} \circ (u \circ v_{i1} \circ \dots \circ v_{i,n_i-1})) + \dots \end{aligned}$$

The “+ . . .” indicate terms containing trees to which we can apply our induction hypothesis. Inside the above expressions, we apply the induction hypothesis to the trees $u \circ v_{i1} \circ \dots \circ v_{i,n_i-1}$, followed once again by Lemma 3.9. We arrive at a huge double sum which constitutes a linear combination of expressions of the form

$$c(u_1 \circ u_2 \circ \dots \circ u_m) \quad (3.30)$$

and of terms “+ . . .” covered by the induction hypothesis. The point of the above dodges was *to make sure that all u_1, u_2, \dots, u_m are in \mathcal{H}* .

Second Step. It remains to reduce an expression (3.30) to the form required by (3.28). The trees u_2, \dots, u_m can be permuted arbitrarily; we arrange them in increasing order $u_2 \leq \dots \leq u_m$.

Case 1. If $u_1 > u_2$, then by definition $u_1 \circ u_2 = w \in \mathcal{H}$ and we absorb the second factor into the first and obtain a product $w \circ u_3 \circ \dots \circ u_m$ with *fewer* factors.

Case 2. If $u_1 < u_2 \leq \dots$, we shuffle the factors with the help of Lemma 3.10 and obtain for (3.30) the expression

$$- \sum_{i=2}^m c(u_i \circ u_1 \circ \dots) + \prod_{i=1}^m c(u_i) + \dots$$

With the first terms we return to Case 1, the second term is precisely as in (3.28), and the terms “+ . . .” are covered by the induction hypothesis.

Case 3. Now let $u_1 = u_2 < \dots$. In this case, the formula (3.25) of Lemma 3.10 contains the term (3.30) twice. We group both together, so that (3.30) becomes

$$- \frac{1}{2} \sum_{i=3}^m c(u_i \circ u_1 \circ u_1 \circ \dots) + \frac{1}{2} \prod_{i=1}^m c(u_i) + \dots$$

and we go back to Case 1. If the first *three* trees are equal, we group three equal terms together and so on.

The whole reduction process is repeated until all Butcher products have disappeared. \square

Theorem 3.14 (Murua & Sanz-Serna 1999). *The composition method $\Psi_h(y) = B_\infty(a_s, y)$ of (3.9) has order p if and only if*

$$a_s(\tau) = \mathbf{e}(\tau) \quad \text{for } \tau \in \mathcal{H} \text{ with } \|\tau\| \leq p.$$

The coefficients $\mathbf{e}(\tau)$ are those of Theorem 3.6.

Proof. We have seen in Sect. II.4 that composition methods of arbitrarily high order exist. Since the coefficients A_i of (3.28) do not depend on the mapping $c(\tau)$, this together with Theorem 3.6 implies that the relation (3.28) is also satisfied by the mapping \mathbf{e} for the exact solution. This proves the statement. \square

Example 3.15. The order conditions for orders $p = 1, \dots, 4$ become, with the trees of Example 3.12 and the rule of (3.14), as follows:

$$\begin{aligned}
\text{Order 1:} \quad & \textcircled{1} \quad \sum_{k=1}^s (\alpha_k + \beta_k) = 1 \\
\text{Order 2:} \quad & \textcircled{2} \quad \sum_{k=1}^s (\alpha_k^2 - \beta_k^2) = 0 \\
\text{Order 3:} \quad & \textcircled{3} \quad \sum_{k=1}^s (\alpha_k^3 + \beta_k^3) = 0 \\
& \textcircled{1} \textcircled{2} \quad \sum_{k=1}^s (\alpha_k^2 - \beta_k^2) \sum_{\ell=1}^k{}' (\alpha_\ell + \beta_\ell) = 0 \\
\text{Order 4:} \quad & \textcircled{4} \quad \sum_{k=1}^s (\alpha_k^4 - \beta_k^4) = 0 \\
& \textcircled{1} \textcircled{3} \quad \sum_{k=1}^s (\alpha_k^3 + \beta_k^3) \sum_{\ell=1}^k{}' (\alpha_\ell + \beta_\ell) = 0 \\
& \textcircled{1} \textcircled{2} \textcircled{1} \quad \sum_{k=1}^s (\alpha_k^2 - \beta_k^2) \left(\sum_{\ell=1}^k{}' (\alpha_\ell + \beta_\ell) \right)^2 = 0,
\end{aligned} \tag{3.31}$$

where, as above, a *prime* attached to a summation symbol indicates that the sum of α_ℓ^i is only from 1 to $k-1$, whereas the sum of $(-\beta_\ell)^i$ is from 1 to k . Similarly, the remaining trees of Example 3.12 with $\|\tau\| = 5$ and $\|\tau\| = 6$ give the additional conditions for order 5 and 6.

We shall see in Sect. V.3 how further reductions and numerical values are obtained under various assumptions of symmetry.

III.3.4 Order Conditions for Splitting Methods

Splitting methods, introduced in Sect. II.5, are based on differential equations of the form

$$\dot{y} = f_1(y) + f_2(y), \tag{3.32}$$

where the flows $\varphi_t^{[1]}$ and $\varphi_t^{[2]}$ of the systems $\dot{y} = f_1(y)$ and $\dot{y} = f_2(y)$ are assumed to be known exactly. In this situation, the method

$$\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]}$$

is of first order and, together with its adjoint $\Phi_h^* = \varphi_h^{[2]} \circ \varphi_h^{[1]}$, can be used as the basic method in the composition (3.9). This yields

$$\Psi_h = \varphi_{a_{s+1}h}^{[1]} \circ \varphi_{b_sh}^{[2]} \circ \varphi_{a_sh}^{[1]} \circ \dots \circ \varphi_{b_2h}^{[2]} \circ \varphi_{a_2h}^{[1]} \circ \varphi_{b_1h}^{[2]} \circ \varphi_{a_1h}^{[1]} \tag{3.33}$$

where

$$b_i = \alpha_i + \beta_i, \quad a_i = \alpha_{i-1} + \beta_i \quad (3.34)$$

with the conventions $\alpha_0 = 0$ and $\beta_{s+1} = 0$. Consequently, the splitting method (3.33) is a special case of (3.9) and we have the following obvious result.

Theorem 3.16. *Suppose that the composition method (3.9) is of order p for all basic methods Φ_h , then the splitting method (3.33) with a_i, b_i given by (3.34) is of the same order p . \square*

We now want to establish the reciprocal result. To every consistent splitting method (3.33), i.e., with coefficients satisfying $\sum_i a_i = \sum_i b_i = 1$, there exist unique α_i, β_i such that (3.34) holds. Does the corresponding composition method have the same order?

Theorem 3.17. *If a consistent splitting method (3.33) is of order p at least for problems of the form (3.32) with the integrable splitting*

$$f_1(y) = \begin{pmatrix} g_1(y_2) \\ 0 \end{pmatrix}, \quad f_2(y) = \begin{pmatrix} 0 \\ g_2(y_1) \end{pmatrix} \quad \text{where} \quad y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad (3.35)$$

then the corresponding composition method has the same order p for an arbitrary basic method Φ_h .

Proof. McLachlan (1995) proves this result in the setting of Lie algebras. We give here a proof using the tools of this section.

a) The flows corresponding to the two vector fields f_1 and f_2 of (3.35) are $\varphi_t^{[1]}(y) = y + tf_1(y)$ and $\varphi_t^{[2]}(y) = y + tf_2(y)$, respectively. Consequently, the method $\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]}$ can be written in the form (3.1) with

$$d_1(y) = f_1(y) + f_2(y), \quad d_{k+1}(y) = \frac{1}{k!} f_1^{(k)}(y) \left(f_2(y), \dots, f_2(y) \right). \quad (3.36)$$

The idea is to construct, for every tree $\tau \in \mathcal{H}$, functions $g_1(y_2)$ and $g_2(y_1)$ such that the first component of $F(\tau)(0)$ is non-zero whereas the first component of $F(\sigma)(0)$ vanishes for all $\sigma \in T_\infty$ different from τ . This construction will be explained in part (b) below. Since the local error of the composition method is a B_∞ -series with coefficients $a_s(\tau) - e(\tau)$, this implies that the order conditions for $\tau \in \mathcal{H}$ with $\|\tau\| \leq p$ are necessary already for this very special class of problems. Theorem 3.14 thus proves the statement.

b) For the construction of the functions $g_1(y_2)$ and $g_2(y_1)$ we have to understand the structure of $F(\tau)(y)$ with $d_k(y)$ given by (3.36). Consider for example the tree $\tau \in T_\infty$ of Fig. 3.2, for which we have $F(\tau)(y) = d_2''(y)(d_1(y), d_3(y))$. Inserting $d_k(y)$ from (3.36), we get by Leibniz' rule a linear combination of eight expressions ($i \in \{1, 2\}$)

$$\begin{aligned} f_1'''(f_2, f_i, f_1''(f_2, f_2)), & \quad f_1''(f_2'f_i, f_1''(f_2, f_2)), \\ f_1''(f_i, f_2'f_1''(f_2, f_2)), & \quad f_1'f_2''(f_i, f_1''(f_2, f_2)), \end{aligned}$$

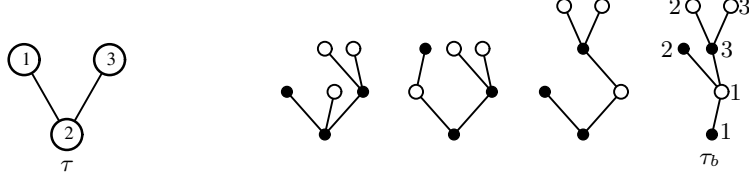


Fig. 3.2. Trees for illustrating the equivalence of the order conditions between composition and splitting methods

each of which can be identified with a bi-coloured tree (see Sect. III.2.1, a vertex \bullet corresponds to f_1 and \circ to f_2). The trees corresponding to these expressions with $i = 1$ are shown in Fig. 3.2. Due to the special form of $d_k(y)$ in (3.36) and due to the fact that in trees of the Hall set \mathcal{H} the vertex ① can appear only at the end of a branch, there is always at least one bi-coloured tree where the vertices \bullet are separated by those of \circ and vice versa. We now select such a tree, denoted by τ_b , and we label the black and white vertices with $\{1, 2, \dots\}$. We then let $y_1 = (y_1^1, \dots, y_n^1)^T$ and $y_2 = (y_1^2, \dots, y_m^2)^T$, where n and m are the numbers of vertices \bullet and \circ in τ_b , respectively. Inspired by “Exercise 4” of Hairer, Nørsett & Wanner (1993), page 155, we define the i th component of $g_1(y_2)$ as the product of all y_j^2 where j runs through the labels of the vertices directly above the vertex \bullet with label i . The function $g_2(y_1)$ is defined similarly. For the example of Fig. 3.2, the tree τ_b yields

$$g_1(y_2) = \begin{pmatrix} y_1^2 \\ y_2^2 y_3^2 \\ 1 \end{pmatrix}, \quad g_2(y_1) = \begin{pmatrix} y_2^1 y_3^1 \\ 1 \\ 1 \end{pmatrix}.$$

One can check that with this construction the bi-coloured tree τ_b is the only one for which the first component of the elementary differential evaluated at $y = 0$ is different from zero. This in turn implies that among all trees of T_∞ only the tree τ has a non-vanishing first component in its elementary differential. \square

Necessity of Negative Steps for Higher Order. One notices that all the composition methods (II.4.6) of order higher than two with Φ_h given by (II.5.7) lead to a splitting (II.5.6) where at least one of the coefficients a_i and b_i is negative. This may be undesirable, especially when the flow $\varphi_t^{[i]}$ originates from a partial differential equation that is ill-posed for negative time progression. The following result has been proved independently by Sheng (1989) and Suzuki (1991) (see also Goldman & Kaper (1996)). We present the elegant proof found by Blanes & Casas (2005).

Theorem 3.18. *If the splitting method (II.5.6) is of order $p \geq 3$ for general $f^{[1]}$ and $f^{[2]}$, then at least one of the a_i and at least one of the b_i are strictly negative.*

Proof. The condition in equation (3.31) for the tree ③ reads

$$\sum_{k=1}^s (\alpha_k^3 + \beta_k^3) = 0 \quad \text{or also} \quad \sum_{k=1}^{s+1} (\alpha_{k-1}^3 + \beta_k^3) = 0$$

(remember that $\alpha_0 = 0$ and $\beta_{s+1} = 0$). Now apply the fact that $x^3 + y^3 < 0$ implies $x + y < 0$ and conclude with formulas (3.34). \square

III.4 The Baker-Campbell-Hausdorff Formula

This section treats the Baker-Campbell-Hausdorff (short BCH or CBH) formula on the composition of exponentials. It was proposed in 1898 by J.E. Campbell and proved independently by Baker (1905) and Hausdorff (1906). This formula will provide an alternative approach to the order conditions of composition (Sect. II.4) and splitting methods (Sect. II.5). For its derivation we shall use the inverse of the derivative of the exponential function.

III.4.1 Derivative of the Exponential and Its Inverse

Elegant formulas for the derivative of \exp and for its inverse can be obtained by the use of matrix commutators $[\Omega, A] = \Omega A - A \Omega$. If we suppose Ω fixed, this expression defines a linear operator $A \mapsto [\Omega, A]$

$$\text{ad}_\Omega(A) = [\Omega, A], \quad (4.1)$$

which is called the adjoint operator (see Varadarajan (1974), Sect. 2.13). Let us start by computing the derivatives of Ω^k . The product rule for differentiation becomes

$$\left(\frac{d}{d\Omega} \Omega^k\right)H = H\Omega^{k-1} + \Omega H\Omega^{k-2} + \dots + \Omega^{k-1}H, \quad (4.2)$$

and this equals $kH\Omega^{k-1}$ if Ω and H commute. Therefore, it is natural to write (4.2) as $kH\Omega^{k-1}$ to which are added correction terms involving commutators and iterated commutators. In the cases $k = 2$ and $k = 3$ we have

$$\begin{aligned} H\Omega + \Omega H &= 2H\Omega + \text{ad}_\Omega(H) \\ H\Omega^2 + \Omega H\Omega + \Omega^2 H &= 3H\Omega^2 + 3(\text{ad}_\Omega(H))\Omega + \text{ad}_\Omega^2(H), \end{aligned}$$

where ad_Ω^i denotes the iterated application of the linear operator ad_Ω . With the convention $\text{ad}_\Omega^0(H) = H$ we obtain by induction on k that

$$\left(\frac{d}{d\Omega} \Omega^k\right)H = \sum_{i=0}^{k-1} \binom{k}{i+1} (\text{ad}_\Omega^i(H))\Omega^{k-i-1}. \quad (4.3)$$

This is seen by applying Leibniz' rule to $\Omega^{k+1} = \Omega \cdot \Omega^k$ and by using the identity $\Omega(\text{ad}_\Omega^i(H)) = (\text{ad}_\Omega^i(H))\Omega + \text{ad}_\Omega^{i+1}(H)$.

Lemma 4.1. *The derivative of $\exp \Omega = \sum_{k \geq 0} \frac{1}{k!} \Omega^k$ is given by*

$$\left(\frac{d}{d\Omega} \exp \Omega\right)H = \left(d \exp_\Omega(H)\right) \exp \Omega,$$

where

$$d \exp_\Omega(H) = \sum_{k \geq 0} \frac{1}{(k+1)!} \text{ad}_\Omega^k(H). \quad (4.4)$$

The series (4.4) converges for all matrices Ω .

Proof. Multiplying (4.3) by $(k!)^{-1}$ and summing, then exchanging the sums and putting $j = k - i - 1$ yields

$$\begin{aligned} \left(\frac{d}{d\Omega} \exp \Omega \right) H &= \sum_{k \geq 0} \frac{1}{k!} \sum_{i=0}^{k-1} \binom{k}{i+1} \left(\text{ad}_{\Omega}^i(H) \right) \Omega^{k-i-1} \\ &= \sum_{i \geq 0} \sum_{j \geq 0} \frac{1}{(i+1)! j!} \left(\text{ad}_{\Omega}^i(H) \right) \Omega^j. \end{aligned}$$

The convergence of the series follows from the boundedness of the linear operator ad_{Ω} (we have $\|\text{ad}_{\Omega}\| \leq 2\|\Omega\|$). \square

Lemma 4.2 (Baker 1905). *If the eigenvalues of the linear operator ad_{Ω} are different from $2\ell\pi i$ with $\ell \in \{\pm 1, \pm 2, \dots\}$, then $d \exp_{\Omega}$ is invertible. Furthermore, we have for $\|\Omega\| < \pi$ that*

$$d \exp_{\Omega}^{-1}(H) = \sum_{k \geq 0} \frac{B_k}{k!} \text{ad}_{\Omega}^k(H), \quad (4.5)$$

where B_k are the Bernoulli numbers, defined by $\sum_{k \geq 0} (B_k/k!) x^k = x/(e^x - 1)$.

Proof. The eigenvalues of $d \exp_{\Omega}$ are $\mu = \sum_{k \geq 0} \lambda^k / (k+1)! = (e^{\lambda} - 1)/\lambda$, where λ is an eigenvalue of ad_{Ω} . By our assumption, the values μ are non-zero, so that $d \exp_{\Omega}$ is invertible. By definition of the Bernoulli numbers, the composition of (4.5) with (4.4) gives the identity. Convergence for $\|\Omega\| < \pi$ follows from $\|\text{ad}_{\Omega}\| \leq 2\|\Omega\|$ and from the fact that the radius of convergence of the series for $x/(e^x - 1)$ is 2π . \square

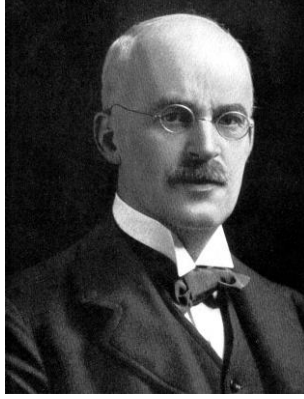
III.4.2 The BCH Formula

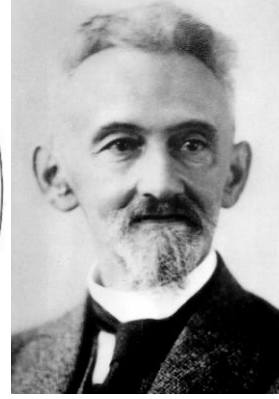
Let A and B be two arbitrary (in general non-commuting) matrices. The problem is to find a matrix $C(t)$, such that

$$\exp(tA) \exp(tB) = \exp C(t). \quad (4.6)$$

In order to get a first idea of the form of $C(t)$, we develop the expression to the left in a series: $\exp(tA) \exp(tB) = I + t(A+B) + \frac{t^2}{2}(A^2 + 2AB + B^2) + \mathcal{O}(t^3) =: I + X$. For sufficiently small t (hence $\|X\|$ is small), the series expansion of the logarithm $\log(I + X) = X - X^2/2 + \dots$ yields a matrix $C(t) = \log(I + X) = t(A+B) + \frac{t^2}{2}(A^2 + 2AB + B^2 - (A+B)^2) + \mathcal{O}(t^3)$, which satisfies (4.6). This series has a positive radius of convergence, because it is obtained by elementary operations of convergent series.

The main problem of the derivation of the BCH formula is to get explicit formulas for the coefficients of the series for $C(t)$, and to express the coefficients of t^2, t^3, \dots in terms of commutators. With the help of the following lemma, recurrence relations for these coefficients will be obtained, which allow for an easy computation of the first terms.


 John Edward Campbell⁴

 Henry Frederick Baker⁵

 Felix Hausdorff⁶

Lemma 4.3. *Let A and B be (non-commuting) matrices. Then, (4.6) holds, where $C(t)$ is the solution of the differential equation*

$$\dot{C} = A + B + \frac{1}{2} [A - B, C] + \sum_{k \geq 2} \frac{B_k}{k!} \text{ad}_C^k(A + B) \quad (4.7)$$

with initial value $C(0) = 0$. Recall that $\text{ad}_C A = [C, A] = CA - AC$, and that B_k denote the Bernoulli numbers as in Lemma 4.2.

Proof. We follow Varadarajan (1974), Sect. 2.15, and we consider for small s and t a smooth matrix function $Z(s, t)$ such that

$$\exp(sA) \exp(tB) = \exp Z(s, t). \quad (4.8)$$

Using Lemma 4.1, the derivative of (4.8) with respect to s is

$$A \exp(sA) \exp(tB) = d \exp_{Z(s,t)} \left(\frac{\partial Z}{\partial s}(s, t) \right) \exp Z(s, t),$$

so that

$$\frac{\partial Z}{\partial s} = d \exp_Z^{-1}(A) = A - \frac{1}{2} [Z, A] + \sum_{k \geq 2} \frac{B_k}{k!} \text{ad}_Z^k(A). \quad (4.9)$$

We next take the inverse of (4.8)

⁴ John Edward Campbell, born: 27 May 1862 in Lisburn, Co Antrim (Ireland), died: 1 October 1924 in Oxford (England).

⁵ Henry Frederick Baker, born: 3 July 1866 in Cambridge (England), died: 17 March 1956 in Cambridge.

⁶ Felix Hausdorff, born: 8 November 1869 in Breslau, Silesia (now Wroclaw, Poland), died: 26 January 1942 in Bonn (Germany).

$$\exp(-tB) \exp(-sA) = \exp(-Z(s, t)),$$

and differentiate this relation with respect to t . As above we get

$$\frac{\partial Z}{\partial t} = d \exp_{-Z}^{-1}(B) = B + \frac{1}{2} [Z, B] + \sum_{k \geq 2} \frac{B_k}{k!} \operatorname{ad}_Z^k(B), \quad (4.10)$$

because $\operatorname{ad}_Z^k(B) = (-1)^k \operatorname{ad}_Z^k(B)$ and the Bernoulli numbers satisfy $B_k = 0$ for odd $k > 2$. A comparison of (4.6) with (4.8) gives $C(t) = Z(t, t)$. The stated differential equation for $C(t)$ therefore follows from $\dot{C}(t) = \frac{\partial Z}{\partial s}(t, t) + \frac{\partial Z}{\partial t}(t, t)$, and from adding the relations (4.9) and (4.10). \square

Using Lemma 4.3 we can compute the first Taylor coefficients of $C(t)$,

$$\exp(tA) \exp(tB) = \exp\left(tC_1 + t^2C_2 + t^3C_3 + t^4C_4 + t^5C_5 + \dots\right). \quad (4.11)$$

Inserting this expansion of $C(t)$ into (4.7) and comparing like powers of t gives

$$\begin{aligned} C_1 &= A + B \\ C_2 &= \frac{1}{4} [A - B, A + B] = \frac{1}{2} [A, B] \\ C_3 &= \frac{1}{6} \left[A - B, \frac{1}{2} [A, B] \right] = \frac{1}{12} [A, [A, B]] + \frac{1}{12} [B, [B, A]] \\ C_4 &= \dots = \frac{1}{24} [A, [B, [B, A]]] \\ C_5 &= \dots = -\frac{1}{720} [A, [A, [A, [A, B]]]] - \frac{1}{720} [B, [B, [B, [B, A]]]] \\ &\quad + \frac{1}{360} [A, [B, [B, [B, A]]]] + \frac{1}{360} [B, [A, [A, [A, B]]]] \\ &\quad + \frac{1}{120} [A, [A, [B, [B, A]]]] + \frac{1}{120} [B, [B, [A, [A, B]]]]. \end{aligned} \quad (4.12)$$

Here, the dots \dots in the formulas for C_4 and C_5 indicate simplifications with the help of the Jacobi identity

$$[A, [B, C]] + [C, [A, B]] + [B, [C, A]] = 0, \quad (4.13)$$

which is verified by straightforward calculation. For higher order the expressions soon become very complicated.

The Symmetric BCH Formula. For the construction of symmetric splitting methods it is convenient to use a formula for the composition

$$\exp\left(\frac{t}{2}A\right) \exp(tB) \exp\left(\frac{t}{2}A\right) = \exp\left(tS_1 + t^3S_3 + t^5S_5 + \dots\right). \quad (4.14)$$

Since the inverse of the left-hand side is obtained by changing the sign of t , the same must be true for the right-hand side. This explains why only odd powers of

t are present in (4.14). Applying the BCH formula (4.11) to $\exp(\frac{t}{2}A)\exp(\frac{t}{2}B) = \exp C(t)$ and a second time to $\exp(C(t))\exp(-C(-t))$ yields for the coefficients of (4.14) (Yoshida 1990)

$$\begin{aligned} S_1 &= A + B \\ S_3 &= -\frac{1}{24} [A, [A, B]] + \frac{1}{12} [B, [B, A]] \\ S_5 &= \frac{7}{5760} [A, [A, [A, [A, B]]]] - \frac{1}{720} [B, [B, [B, [B, A]]]] \\ &\quad + \frac{1}{360} [A, [B, [B, [B, A]]]] + \frac{1}{360} [B, [A, [A, [A, B]]]] \\ &\quad - \frac{1}{480} [A, [A, [B, [B, A]]]] + \frac{1}{120} [B, [B, [A, [A, B]]]]. \end{aligned} \quad (4.15)$$

III.5 Order Conditions via the BCH Formula

Using the BCH formula we present an alternative approach to the order conditions of splitting and composition methods. The main idea is to write the flow of a differential equation formally as the exponential of the Lie derivative.

III.5.1 Calculus of Lie Derivatives

For a differential equation

$$\dot{y} = f^{[1]}(y) + f^{[2]}(y),$$

it is convenient to study the composition of the flows $\varphi_t^{[1]}$ and $\varphi_t^{[2]}$ of the systems

$$\dot{y} = f^{[1]}(y), \quad \dot{y} = f^{[2]}(y), \quad (5.1)$$

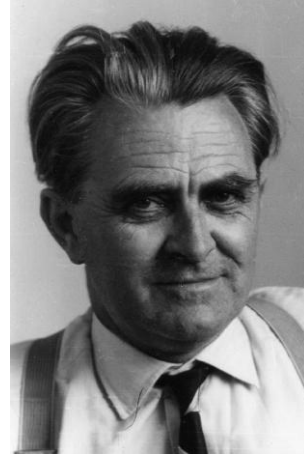
respectively. We introduce the differential operators (*Lie derivative*)

$$D_i = \sum_j f_j^{[i]}(y) \frac{\partial}{\partial y_j}$$

which means that for differentiable functions $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ we have

$$D_i F(y) = F'(y) f^{[i]}(y). \quad (5.2)$$

It follows from the chain rule that, for the solutions $\varphi_t^{[i]}(y_0)$ of (5.1),



Wolfgang Gröbner⁷

⁷ Wolfgang Gröbner, born: 11 February 1899 in Gossensass, South Tyrol (now Italy), died: 10 August 1980 in Innsbruck.

$$\frac{d}{dt} F(\varphi_t^{[i]}(y_0)) = (D_i F)(\varphi_t^{[i]}(y_0)), \quad (5.3)$$

and applying this operator iteratively we get

$$\frac{d^k}{dt^k} F(\varphi_t^{[i]}(y_0)) = (D_i^k F)(\varphi_t^{[i]}(y_0)). \quad (5.4)$$

Consequently, the Taylor series of $F(\varphi_t^{[i]}(y_0))$, developed at $t = 0$, becomes

$$F(\varphi_t^{[i]}(y_0)) = \sum_{k \geq 0} \frac{t^k}{k!} (D_i^k F)(y_0) = \exp(tD_i)F(y_0). \quad (5.5)$$

Now, putting $F(y) = \text{Id}(y) = y$, the identity map, this is the Taylor series of the solution itself

$$\varphi_t^{[i]}(y_0) = \sum_{k \geq 0} \frac{t^k}{k!} (D_i^k \text{Id})(y_0) = \exp(tD_i)\text{Id}(y_0). \quad (5.6)$$

If the functions $f^{[i]}(y)$ are not analytic, but only N -times continuously differentiable, the series (5.6) has to be truncated and a $\mathcal{O}(h^N)$ remainder term has to be included.

Lemma 5.1 (Gröbner 1960). *Let $\varphi_s^{[1]}$ and $\varphi_t^{[2]}$ be the flows of the differential equations $\dot{y} = f^{[1]}(y)$ and $\dot{y} = f^{[2]}(y)$, respectively. For their composition we then have*

$$(\varphi_t^{[2]} \circ \varphi_s^{[1]})(y_0) = \exp(sD_1) \exp(tD_2) \text{Id}(y_0).$$

Proof. This is precisely formula (5.5) with $i = 1$, t replaced with s , and with $F(y) = \varphi_t^{[2]}(y) = \exp(tD_2)\text{Id}(y_0)$. \square

Remark 5.2. Notice that the indices 1 and 2 as well as s and t to the left and right in the identity of Lemma 5.1 are permuted. Gröbner calls this phenomenon, which sometimes leads to some confusion in the literature, the “Vertauschungssatz”.

Remark 5.3. The statement of Lemma 5.1 can be extended to more than two flows. If $\varphi_t^{[j]}$ is the flow of a differential equation $\dot{y} = f^{[j]}(y)$, then we have

$$(\varphi_u^{[m]} \circ \dots \circ \varphi_t^{[2]} \circ \varphi_s^{[1]})(y_0) = \exp(sD_1) \exp(tD_2) \dots \exp(uD_m) \text{Id}(y_0).$$

This follows by induction on m .

In general, the two operators D_1 and D_2 do not commute, so that the composition $\exp(tD_1) \exp(tD_2)\text{Id}(y_0)$ is different from $\exp(t(D_1 + D_2))\text{Id}(y_0)$, which represents the solution $\varphi_t(y_0)$ of $\dot{y} = f(y) = f^{[1]}(y) + f^{[2]}(y)$. The relation of Lemma 5.1 suggests the use of the BCH formula. However, D_1 and D_2 are unbounded differential operators so that the series expansions that appear cannot be

expected to converge. A formal application of the BCH formula with tA and tB replaced with sD_1 and tD_2 , respectively, yields

$$\exp(sD_1)\exp(tD_2) = \exp(D(s, t)), \quad (5.7)$$

where the differential operator $D(s, t)$ is obtained from (4.11) as

$$\begin{aligned} D(s, t) = & sD_1 + tD_2 + \frac{st}{2}[D_1, D_2] + \frac{s^2t}{12}[D_1, [D_1, D_2]] \\ & + \frac{st^2}{12}[D_2, [D_2, D_1]] + \frac{s^2t^2}{24}[D_1, [D_2, [D_2, D_1]]] + \dots \end{aligned} \quad (5.8)$$

The *Lie bracket* for differential operators is calculated exactly as for matrices, namely, $[D_1, D_2] = D_1D_2 - D_2D_1$. But how can we interpret (5.7) rigorously? Expanding both sides in Taylor series we see that

$$\exp(sD_1)\exp(tD_2) = I + sD_1 + tD_2 + \frac{1}{2}(s^2D_1^2 + 2stD_1D_2 + t^2D_2^2) + \dots \quad (5.9)$$

and

$$\begin{aligned} \exp(D(s, t)) &= I + D(s, t) + \frac{1}{2}D(s, t)^2 + \dots \\ &= I + sD_1 + tD_2 + \frac{1}{2}((sD_1 + tD_2)^2 + st[D_1, D_2]) + \dots \end{aligned}$$

By derivation of the BCH formula we have a formal identity, i.e., both series have exactly the same coefficients. Moreover, every finite truncation of the series can be applied without any difficulties to sufficiently differentiable functions $F(y)$. Consequently, for N -times differentiable functions the relation (5.7) holds true, if both sides are replaced by their truncated Taylor series and if a $\mathcal{O}(h^N)$ remainder is added ($h = \max(|s|, |t|)$).

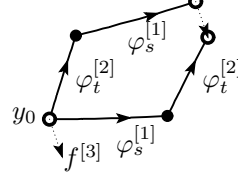
III.5.2 Lie Brackets and Commutativity

If we apply D_2 to a function F , followed by an application of D_1 , we will obtain partial derivatives of F of first and second orders. However, if we subtract from this the same expression with D_1 and D_2 reversed, the second derivatives will cancel (this was already remarked upon by Jacobi (1862), p. 39: “differentialia partialia secunda functionis f non continere”) and we see that the Lie bracket

$$[D_1, D_2] = D_1D_2 - D_2D_1 = \sum_i \left(\sum_j \left(\frac{\partial f_i^{[2]}}{\partial y_j} f_j^{[1]} - \frac{\partial f_i^{[1]}}{\partial y_j} f_j^{[2]} \right) \right) \frac{\partial}{\partial y_i} \quad (5.10)$$

is again a linear differential operator. So, from two vector fields $f^{[1]}$ and $f^{[2]}$ we obtain a *third* vector field $f^{[3]}$.

The *geometric meaning* of the new vector field can be deduced from Lemma 5.1. We see by subtracting (5.9) from itself, once as it stands and once with sD_1 and tD_2 permuted, that



$$\varphi_t^{[2]} \circ \varphi_s^{[1]}(y_0) - \varphi_s^{[1]} \circ \varphi_t^{[2]}(y_0) = st [D_1, D_2] \text{Id}(y_0) + \dots = st f^{[3]}(y_0) + \dots \quad (5.11)$$

(see the picture), where “+ ...” are terms of order ≥ 3 . This leads us to the following result.

Lemma 5.4. *Let $f^{[1]}(y)$ and $f^{[2]}(y)$ be defined on an open set. The corresponding flows $\varphi_s^{[1]}$ and $\varphi_t^{[2]}$ commute everywhere for all sufficiently small s and t , if and only if*

$$[D_1, D_2] = 0. \quad (5.12)$$

Proof. The “only if” part is clear from (5.11). For proving the “if” part, we take s and t fixed, and subdivide, for a given n , the integration intervals into n equidistant parts $\Delta s = s/n$ and $\Delta t = t/n$. This allows us to transform the solution $\varphi_t^{[2]} \circ \varphi_s^{[1]}(y_0)$ by a discrete homotopy in n^2 steps into the solution $\varphi_s^{[1]} \circ \varphi_t^{[2]}(y_0)$, each time appending a small rectangle of size $\mathcal{O}(n^{-2})$. If we denote such an intermediate stage by

$$\Gamma_k = \dots \circ \varphi_{j_2 \Delta t}^{[2]} \circ \varphi_{i_2 \Delta s}^{[1]} \circ \varphi_{j_1 \Delta t}^{[2]} \circ \varphi_{i_1 \Delta s}^{[1]}(y_0)$$

then we have $\Gamma_0 = \varphi_t^{[2]} \circ \varphi_s^{[1]}(y_0)$ and $\Gamma_{n^2} = \varphi_s^{[1]} \circ \varphi_t^{[2]}(y_0)$ (see Fig. 5.1). Now, for $n \rightarrow \infty$, we have the estimate

$$|\Gamma_{k+1} - \Gamma_k| \leq \mathcal{O}(n^{-3}),$$

because the error terms in (5.11) are of order 3 at least, and because of the differentiability of the solutions with respect to initial values. Thus, by the triangle inequality $|\Gamma_{n^2} - \Gamma_0| \leq \mathcal{O}(n^{-1})$ and the result is proved. \square

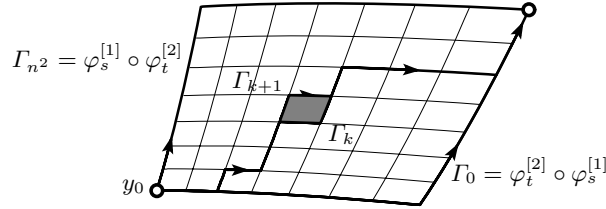


Fig. 5.1. Estimation of commuting solutions

III.5.3 Splitting Methods

We follow the approach of Yoshida (1990) for obtaining the order conditions of splitting methods (II.5.6). The idea is the following: with the use of Lemma 5.1 we write the method as a product of exponentials, then we apply formally the Baker-Campbell-Hausdorff formula to get one exponential of a series in powers of h . Finally, we compare this series with $h(D_1 + D_2)$, which corresponds to the exact solution of (5.1).

The splitting method (II.5.6), viz.,

$$\Psi_h = \varphi_{b_m h}^{[2]} \circ \varphi_{a_m h}^{[1]} \circ \varphi_{b_{m-1} h}^{[2]} \circ \dots \circ \varphi_{a_2 h}^{[1]} \circ \varphi_{b_1 h}^{[2]} \circ \varphi_{a_1 h}^{[1]}, \quad (5.13)$$

is a composition of expressions $\varphi_{b_j h}^{[2]} \circ \varphi_{a_j h}^{[1]}$ which, by Lemma 5.1 and by (5.7), can be written as an exponential

$$\begin{aligned} \varphi_{b_j h}^{[2]} \circ \varphi_{a_j h}^{[1]} = \exp \Big(& a_j h E_1^1 + b_j h E_2^1 + a_j b_j h^2 E_1^2 \\ & + a_j^2 b_j h^3 E_1^3 + a_j b_j^2 h^3 E_2^3 + a_j^2 b_j^2 h^4 E_1^4 + \dots \Big) \text{Id}, \end{aligned} \quad (5.14)$$

where we use the abbreviations

$$\begin{aligned} E_1^1 &= D_1, & E_2^1 &= D_2, & E_1^2 &= \frac{1}{2}[D_1, D_2], & E_1^3 &= \frac{1}{12}[D_1, [D_1, D_2]], \\ E_2^3 &= \frac{1}{12}[D_2, [D_2, D_1]], & E_1^4 &= \frac{1}{24}[D_1, [D_2, [D_2, D_1]]], \end{aligned}$$

and the dots indicate $\mathcal{O}(h^5)$ expressions.

We next define $\Psi^{(j)}$ recursively by

$$\Psi^{(0)} = \text{Id}, \quad \Psi^{(j)} = \varphi_{b_j h}^{[2]} \circ \varphi_{a_j h}^{[1]} \circ \Psi^{(j-1)}, \quad (5.15)$$

so that $\Psi^{(m)}$ is equal to our method (5.13). Aiming to write $\Psi^{(j)}$ also as an exponential of differential operators, we are confronted with computing commutators of the expressions E_i^j . We see that $[E_1^1, E_2^1] = 2E_1^2$, $[E_1^1, E_1^2] = 6E_1^3$, $[E_2^1, E_1^2] = -6E_2^3$, $[E_1^1, E_2^3] = 2E_1^4$, and $[E_2^1, E_1^3] = -2E_1^4$ as a consequence of the Jacobi identity (4.13). But the other commutators cannot be expressed in terms of E_i^j . We therefore introduce

$$E_2^4 = \frac{1}{24}[D_1, [D_1, [D_1, D_2]]], \quad E_3^4 = \frac{1}{24}[D_2, [D_2, [D_2, D_1]]].$$

This allows us to formulate the following result.

Lemma 5.5. *The method $\Psi^{(j)}$, defined by (5.15), can be formally written as*

$$\begin{aligned} \Psi^{(j)} = \exp \Big(& c_{1,j}^1 h E_1^1 + c_{2,j}^1 h E_2^1 + c_{1,j}^2 h^2 E_1^2 + c_{1,j}^3 h^3 E_1^3 \\ & + c_{2,j}^3 h^3 E_2^3 + c_{1,j}^4 h^4 E_1^4 + c_{2,j}^4 h^4 E_2^4 + c_{3,j}^4 h^4 E_3^4 + \dots \Big) \text{Id}, \end{aligned}$$

where all coefficients are zero for $j = 0$, and where for $j \geq 1$

$$\begin{aligned} c_{1,j}^1 &= c_{1,j-1}^1 + a_j, & c_{2,j}^1 &= c_{2,j-1}^1 + b_j, \\ c_{1,j}^2 &= c_{1,j-1}^2 + a_j b_j + c_{1,j-1}^1 b_j - c_{2,j-1}^1 a_j, \\ c_{1,j}^3 &= c_{1,j-1}^3 + a_j^2 b_j + 2c_{1,j-1}^1 a_j b_j - 3c_{1,j-1}^2 a_j \\ &\quad + (c_{1,j-1}^1)^2 b_j - c_{1,j-1}^1 c_{2,j-1}^1 a_j + c_{2,j-1}^1 a_j^2, \\ c_{2,j}^3 &= c_{2,j-1}^3 + a_j b_j^2 - 4c_{2,j-1}^1 a_j b_j + 3c_{1,j-1}^2 b_j \\ &\quad + (c_{2,j-1}^1)^2 a_j - c_{1,j-1}^1 c_{2,j-1}^1 b_j + c_{1,j-1}^1 b_j^2, \end{aligned}$$

and similar but more complicated formulas for $c_{i,j}^4$.

Proof. Due to the reversed order in Lemma 5.1 we have to compute $\exp(A) \exp(B)$, where A is the argument of the exponential for $\Psi^{(j-1)}$ and B is that of (5.14). The rest is a tedious but straightforward application of the BCH formula. One has to use repeatedly the formulas for $[E_i^j, E_k^l]$, stated before Lemma 5.5. \square

Theorem 5.6. *The splitting method (5.13) is of order p if*

$$c_{1,m}^1 = c_{2,m}^1 = 1, \quad c_{\ell,m}^k = 0 \quad \text{for } k = 2, \dots, p \text{ and all } \ell. \quad (5.16)$$

The coefficients $c_{\ell,m}^k$ are those defined in Lemma 5.5.

Proof. This is an immediate consequence of Lemma 5.5, because the conditions of order p imply that the Taylor series expansion of $\Psi^{(m)}(y_0)$ coincides with that of the solution $\varphi_h(y_0) = \exp(h(D_1 + D_2))y_0$ up to terms of size $\mathcal{O}(h^p)$. \square

A simplification in the order conditions arises for symmetric methods (5.13), that is, for coefficients satisfying $a_{m+1-i} = a_i$ and $b_{m-i} = b_i$ for all i (and $b_m = 0$). By Theorem II.3.2, it is sufficient to consider the order conditions (5.16) for odd k only.

III.5.4 Composition Methods

We now consider composition methods (II.4.6), viz.,

$$\Psi_h = \Phi_{\alpha_s h} \circ \Phi_{\beta_s h}^* \circ \dots \circ \Phi_{\beta_2 h}^* \circ \Phi_{\alpha_1 h} \circ \Phi_{\beta_1 h}^*, \quad (5.17)$$

where Φ_h is a first-order method for $\dot{y} = f(y)$ and Φ_h^* is its adjoint. We assume

$$\Phi_h = \exp\left(hC_1 + h^2C_2 + h^3C_3 + \dots\right) \text{Id} \quad (5.18)$$

with differential operators C_i , and such that C_1 is the Lie derivative operator corresponding to $\dot{y} = f(y)$. For the splitting method $\Phi_h = \varphi_h^{[2]} \circ \varphi_h^{[1]}$ this follows from (5.14), and for general one-step methods this is a consequence of Sect. IX.1 on backward error analysis. The adjoint method then satisfies

$$\Phi_h^* = \exp(hC_1 - h^2C_2 + h^3C_3 - \dots)\text{Id}. \quad (5.19)$$

From now on the procedure is similar to that of Sect. III.5.3. We define $\Psi^{(j)}$ recursively by

$$\Psi^{(0)} = \text{Id}, \quad \Psi^{(j)} = \Phi_{\alpha_j h} \circ \Phi_{\beta_j h}^* \circ \Psi^{(j-1)}, \quad (5.20)$$

so that $\Psi^{(m)}$ becomes (5.17). We apply the BCH formula to obtain

$$\begin{aligned} \Phi_{\alpha_j h} \circ \Phi_{\beta_j h}^* &= \exp(\beta_j h C_1 - \beta_j^2 h^2 C_2 + \dots) \exp(\alpha_j h C_1 + \alpha_j^2 h^2 C_2 + \dots) \text{Id} \\ &= \exp\left((\alpha_j + \beta_j) h E_1^1 + (\alpha_j^2 - \beta_j^2) h^2 E_1^2 \right. \\ &\quad \left. + (\alpha_j^3 + \beta_j^3) h^3 E_1^3 + \frac{1}{2} \alpha_j \beta_j (\alpha_j + \beta_j) h^3 E_2^3 + \dots\right) \text{Id} \end{aligned}$$

where

$$E_1^k = C_k, \quad E_2^3 = [C_1, C_2].$$

We then have the following result.

Lemma 5.7. *The method $\Psi^{(j)}$ of (5.20) can be formally written as*

$$\Psi^{(j)} = \exp\left(\gamma_{1,j}^1 h E_1^1 + \gamma_{1,j}^2 h^2 E_1^2 + \gamma_{1,j}^3 h^3 E_1^3 + \gamma_{2,j}^3 h^3 E_2^3 + \dots\right) \text{Id},$$

where all coefficients are zero for $j = 0$, and where for $j = 1, \dots, m$

$$\begin{aligned} \gamma_{1,j}^1 &= \gamma_{1,j-1}^1 + \alpha_j + \beta_j \\ \gamma_{1,j}^2 &= \gamma_{1,j-1}^2 + \alpha_j^2 - \beta_j^2 \\ \gamma_{1,j}^3 &= \gamma_{1,j-1}^3 + \alpha_j^3 + \beta_j^3 \\ \gamma_{2,j}^3 &= \gamma_{2,j-1}^3 + \frac{1}{2} \alpha_j \beta_j (\alpha_j + \beta_j) + \frac{1}{2} \gamma_{1,j-1}^1 (\alpha_j^2 - \beta_j^2) - \frac{1}{2} \gamma_{1,j-1}^2 (\alpha_j + \beta_j). \end{aligned}$$

Proof. Similar to Lemma 5.5, the result follows using the BCH formula. \square

Theorem 5.8. *The composition method (5.17) is of order p if*

$$\gamma_{1,m}^1 = 1, \quad \gamma_{\ell,m}^k = 0 \quad \text{for } k = 2, \dots, p \text{ and all } \ell. \quad (5.21)$$

The coefficients $\gamma_{\ell,m}^k$ are those defined in Lemma 5.7. \square

It is interesting to see how these order conditions are related to those obtained with the use of trees. The conditions $\gamma_{1,m}^1 = 1$ and $\gamma_{1,m}^2 = \gamma_{1,m}^3 = 0$ are identical to the first three order conditions of Example 3.15. The remaining condition for order 3, $\gamma_{2,m}^3 = 0$, reads

$$\begin{aligned} &\sum_{k=1}^m \alpha_k \beta_k (\alpha_k + \beta_k) + \sum_{k=1}^m (\alpha_k^2 - \beta_k^2) \sum_{i=1}^{k-1} (\alpha_i + \beta_i) - \sum_{k=1}^m (\alpha_k + \beta_k) \sum_{i=1}^{k-1} (\alpha_i^2 - \beta_i^2) \\ &= \sum_{k=1}^m (\alpha_k^2 - \beta_k^2) \sum_{i=1}^k (\alpha_i + \beta_i) - \sum_{k=1}^m (\alpha_k + \beta_k) \sum_{i=1}^k (\alpha_i^2 - \beta_i^2) = 0. \end{aligned}$$

This condition is just the difference of the order conditions for the trees $\textcircled{2} \circ \textcircled{1}$ and $\textcircled{1} \circ \textcircled{2}$, whose sum is zero by the Switching Lemma 3.8. Therefore the condition $\gamma_{2,m}^3 = 0$ is equivalent to (though more complicated than) the fourth condition of Example 3.15.

Symmetric Composition of Symmetric Methods. Consider now a composition

$$\Psi_h = \Phi_{\gamma_m h} \circ \dots \circ \Phi_{\gamma_2 h} \circ \Phi_{\gamma_1 h} \circ \Phi_{\gamma_2 h} \circ \dots \circ \Phi_{\gamma_m h}, \quad (5.22)$$

where Φ_h is a symmetric method that can be written as

$$\Phi_h = \exp(hS_1 + h^3S_3 + h^5S_5 + \dots) \text{Id}$$

with S_1 the Lie derivative operator corresponding to $\dot{y} = f(y)$. For the Strang splitting $\Phi_h = \varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}$ such an expansion follows from the symmetric BCH formula (4.14), and for general symmetric one-step methods from Sect. IX.2. The derivation of the order conditions is similar to the above with $\Psi^{(j)}$ defined by

$$\Psi^{(1)} = \Phi_{\gamma_1 h}, \quad \Psi^{(j)} = \Phi_{\gamma_j h} \circ \Psi^{(j-1)} \circ \Phi_{\gamma_j h},$$

so that $\Psi^{(m)}$ becomes (5.22).

Lemma 5.9. *The method $\Psi^{(j)}$ can be formally written as*

$$\Psi^{(j)} = \exp(\sigma_{1,j}^1 h E_1^1 + \sigma_{1,j}^3 h^3 E_1^3 + \sigma_{1,j}^5 h^5 E_1^5 + \sigma_{2,j}^5 h^5 E_2^5 + \dots) \text{Id},$$

where $E_1^k = S_k$, $E_2^5 = [S_1[S_1, S_3]]$, and where $\sigma_{1,1}^k = \gamma_1^k$, $\sigma_{2,1}^5 = 0$, and

$$\begin{aligned} \sigma_{1,j}^k &= \sigma_{1,j-1}^k + 2\gamma_j^k \\ \sigma_{2,j}^5 &= \sigma_{2,j-1}^5 + \frac{1}{6} \left(\gamma_j^3 (\sigma_{1,j-1}^1)^2 - \gamma_j \sigma_{1,j-1}^1 \sigma_{1,j-1}^3 - \gamma_j^2 \sigma_{1,j-1}^3 + \gamma_j^4 \sigma_{1,j-1}^1 \right). \end{aligned}$$

Proof. The result is a consequence of the symmetric BCH formula (4.14) with $\gamma_j h S_1 + \gamma_j^3 h^3 S_3 + \dots$ and $\sigma_{1,j-1}^1 h E_1^1 + \sigma_{1,j-1}^3 h E_1^3 + \dots$ in the roles of $\frac{t}{2}A$ and tB , respectively. \square

Theorem 5.10. *The composition method (5.22) is of order p if*

$$\sigma_{1,m}^1 = 1, \quad \sigma_{\ell,m}^k = 0 \quad \text{for odd } k = 3, \dots, p \text{ and all } \ell. \quad (5.23)$$

The coefficients $\sigma_{\ell,m}^k$ are those defined in Lemma 5.9. \square

Symmetric composition methods up to order 10 will be constructed and discussed in Sect. V.3.

III.6 Exercises

- Find all trees of orders 5 and 6.
- (A. Cayley 1857). Denote the number of trees of order q by a_q . Prove that

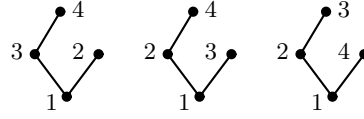
$$a_1 + a_2x + a_3x^2 + a_4x^3 + \dots = (1-x)^{-a_1}(1-x^2)^{-a_2}(1-x^3)^{-a_3} \dots$$

q	1	2	3	4	5	6	7	8	9	10
a_q	1	1	2	4	9	20	48	115	286	719

- Independency of the elementary differentials: show that for every $\tau \in T$ there is a system (1.1) such that the first component of $F(\tau)(0)$ equals 1, and the first component of $F(u)(0)$ is zero for all trees $u \neq \tau$.

Hint. Consider a monotonic labelling of τ , and define y'_i as the product over all y_j , where j runs through all labels of vertices that lie directly above the vertex “ i ”. For the first labelling of the tree of Exercise 4 this would be $y'_1 = y_2y_3$, $y'_2 = 1$, $y'_3 = y_4$, and $y'_4 = 1$.

- Prove that the coefficient $\alpha(\tau)$ of Definition 1.2 is equal to the number of possible monotonic labellings of the vertices of τ , starting with the label 1 for the root. For example, the tree $[[\bullet], \bullet]$ has three different monotonic labellings.



In addition, deduce, from (1.22), the recursion formula

$$\alpha(\tau) = \binom{|\tau| - 1}{|\tau_1|, \dots, |\tau_m|} \alpha(\tau_1) \dots \alpha(\tau_m) \frac{1}{\mu_1! \mu_2! \dots}, \quad (6.1)$$

where the integers μ_1, μ_2, \dots count equal trees among τ_1, \dots, τ_m and

$$\binom{|\tau| - 1}{|\tau_1|, \dots, |\tau_m|} = \frac{(|\tau| - 1)!}{|\tau_1|! \dots |\tau_m|!}$$

denotes the multinomial coefficient.

Remark. In the theoretical physics literature, the coefficients $\alpha(\tau)$ are written $CM(\tau)$ and called “Connes-Moscovici weights”.

- If we denote by $N(\tau)$ the number of elements in $OST(\tau)$, then show that

$$N(\bullet) = 2, \quad N([\tau_1, \dots, \tau_m]) = 1 + N(\tau_1) \dots N(\tau_m).$$

Use this result to compute the number of subtrees of the christmas tree decorating formula (1.34). *Answer:* 6865.

- Prove that the elementary differentials for partitioned problems are independent. For a given tree ($\tau \in TP$), find a problem (2.1) such that a certain component of $F(\tau)(p, q)$ vanishes for all $u \in TP$ except for τ .

Hint. Consider the construction of Exercise 3, and define the partitioning of y into (p, q) according to the colours of the vertices.

7. The number of order conditions for partitioned Runge–Kutta methods (II.2.2) is $2a_r$ for order r , where a_r is given by (see Hairer, Nørsett & Wanner (1993), page 311)

r	1	2	3	4	5	6	7	8	9	10
a_r	1	2	7	26	107	458	2058	9498	44987	216598

Find a formula similar to that of Exercise 2.

8. For the special second order differential equation $\ddot{y} = g(y)$, and for a Nyström method

$$\begin{aligned}\ell_i &= g\left(y_0 + c_i h \dot{y}_0 + h^2 \sum_{j=1}^s a_{ij} \ell_j\right), \\ y_1 &= y_0 + h \dot{y}_0 + h^2 \sum_{i=1}^s \beta_i \ell_i, \quad \dot{y}_1 = \dot{y}_0 + h \sum_{i=1}^s b_i \ell_i,\end{aligned}\tag{6.2}$$

consider the simplifying assumption

$$\begin{aligned}CN(\eta) : \quad & \sum_{j=1}^s a_{ij} c_j^{k-2} = \frac{c_i^k}{k(k-1)}, \quad k = 2, \dots, \eta, \\ DN(\zeta) : \quad & \sum_{i=1}^s b_i c_i^{k-2} a_{ij} = b_j \left(\frac{c_j^k}{k(k-1)} - \frac{c_j}{k-1} + \frac{1}{k} \right), \quad k = 2, \dots, \zeta.\end{aligned}$$

Prove that if the quadrature formula (b_i, c_i) is of order p , if $\beta_i = b_i(1 - c_i)$ for all i , and if the simplifying assumptions $CN(\eta)$, $DN(\zeta)$ are satisfied with $2\eta + 2 \geq p$ and $\zeta + \eta \geq p$, then the Nyström method has order p .

9. *Nyström methods of maximal order $2s$.* Prove that there exists a one-parameter family of s -stage Nyström methods (6.2) for $\ddot{y} = g(y)$, which have order $2s$.
Hint. Consider the Gaussian quadrature formula and define the coefficients a_{ij} by $CN(s)$ and by

$$\sum_{i=1}^s b_i c_i^{k-2} a_{is} = b_s \left(\frac{c_s^k}{k(k-1)} - \frac{c_s}{k-1} + \frac{1}{k} \right)$$

for $k = 2, \dots, s$.

10. Prove that the coefficient C_4 in the series (4.11) of the Baker–Campbell–Hausdorff formula is given by $C_4 = [A, [B, [B, A]]]/24$.
11. Prove that the series (4.11) converges for $|t| < \ln 2/(\|A\| + \|B\|)$.
12. By Theorem 5.10 four order conditions have to be satisfied such that the symmetric composition method (5.22) is of order 6. Prove that these conditions are equivalent to the four conditions of Example V.3.15. (Care has to be taken due to the different meaning of the γ_i .)

Chapter IV.

Conservation of First Integrals and Methods on Manifolds

This chapter deals with the conservation of invariants (first integrals) by numerical methods, and with numerical methods for differential equations on manifolds. Our investigation will follow two directions. We first investigate which of the methods introduced in Chap. II conserve invariants automatically. We shall see that most of them conserve linear invariants, a few of them quadratic invariants, and none of them conserves cubic or general nonlinear invariants. We then construct new classes of methods, which are adapted to known invariants and which force the numerical solution to satisfy them. In particular, we study projection methods and methods based on local coordinates of the manifold defined by the invariants. We discuss in some detail the case where the manifold is a Lie group. Finally, we consider differential equations on manifolds with orthogonality constraints, which often arise in numerical linear algebra.

IV.1 Examples of First Integrals

Je nomme intégrale une équation $u = \text{Const.}$ telle que sa différentielle $du = 0$ soit vérifiée identiquement par le système des équations différentielles proposées . . . (C.G.J. Jacobi 1840, p. 350)

We consider differential equations

$$\dot{y} = f(y), \quad (1.1)$$

where y is a vector or possibly a matrix.

Definition 1.1. A non-constant function $I(y)$ is called a *first integral* of (1.1) if

$$I'(y)f(y) = 0 \quad \text{for all } y. \quad (1.2)$$

This implies that *every* solution $y(t)$ of (1.1) satisfies $I(y(t)) = I(y_0) = \text{Const.}$ Synonymously with “first integral”, the terms *invariant* or *conserved quantity* or *constant of motion* are also used.

In Chap. I we have seen many examples of differential equations with invariants. For example, the Lotka–Volterra problem (I.1.1) has $I(u, v) = \ln u - u + 2 \ln v - v$ as first integral. The pendulum equation (I.1.13) has $H(p, q) = p^2/2 - \cos q$, and the Kepler problem (I.2.2) has two first integrals, namely H and L of (I.2.3) and (I.2.4).

Example 1.2 (Conservation of the Total Energy). Hamiltonian systems are of the form

$$\dot{p} = -H_q(p, q), \quad \dot{q} = H_p(p, q),$$

where $H_q = \nabla_q H = (\partial H / \partial q)^T$ and $H_p = \nabla_p H = (\partial H / \partial p)^T$ are the column vectors of partial derivatives. The Hamiltonian function $H(p, q)$ is a first integral. This follows at once from $H'(p, q) = (\partial H / \partial p, \partial H / \partial q)$ and

$$\frac{\partial H}{\partial p} \left(-\frac{\partial H}{\partial q} \right)^T + \frac{\partial H}{\partial q} \left(\frac{\partial H}{\partial p} \right)^T = 0.$$

Example 1.3 (Conservation of the Total Linear and Angular Momentum of N-Body Systems). We consider a system of N particles interacting pairwise with potential forces which depend on the distances of the particles. This is formulated as a Hamiltonian system with total energy (I.4.1), viz.,

$$H(p, q) = \frac{1}{2} \sum_{i=1}^N \frac{1}{m_i} p_i^T p_i + \sum_{i=2}^N \sum_{j=1}^{i-1} V_{ij}(\|q_i - q_j\|).$$

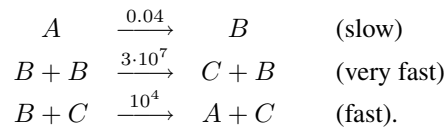
Here $q_i, p_i \in \mathbb{R}^3$ represent the position and momentum of the i th particle of mass m_i , and $V_{ij}(r)$ ($i > j$) is the interaction potential between the i th and j th particle. The equations of motion read

$$\dot{q}_i = \frac{1}{m_i} p_i, \quad \dot{p}_i = \sum_{j=1}^N \nu_{ij} (q_i - q_j)$$

where, for $i > j$, we have $\nu_{ij} = \nu_{ji} = -V'_{ij}(r_{ij})/r_{ij}$ with $r_{ij} = \|q_i - q_j\|$, and ν_{ii} is arbitrary, say $\nu_{ii} = 0$. The conservation of the total *linear momentum* $P = \sum_{i=1}^N p_i$ and the *angular momentum* $L = \sum_{i=1}^N q_i \times p_i$ is a consequence of the symmetry relation $\nu_{ij} = \nu_{ji}$:

$$\begin{aligned} \frac{d}{dt} \sum_{i=1}^N p_i &= \sum_{i=1}^N \sum_{j=1}^N \nu_{ij} (q_i - q_j) = 0 \\ \frac{d}{dt} \sum_{i=1}^N q_i \times p_i &= \sum_{i=1}^N \frac{1}{m_i} p_i \times p_i + \sum_{i=1}^N \sum_{j=1}^N q_i \times \nu_{ij} (q_i - q_j) = 0. \end{aligned}$$

Example 1.4 (Conservation of Mass in Chemical Reactions). Suppose that three substances A, B, C undergo a chemical reaction such as¹



¹ This *Robertson problem* is very popular in testing codes for stiff differential equations.

We denote the masses (or concentrations) of the substances A, B, C by y_1, y_2, y_3 , respectively. By the mass action law this leads to the equations

$$\begin{aligned} \text{A:} \quad & \dot{y}_1 = -0.04 y_1 + 10^4 y_2 y_3 \\ \text{B:} \quad & \dot{y}_2 = 0.04 y_1 - 10^4 y_2 y_3 - 3 \cdot 10^7 y_2^2 \\ \text{C:} \quad & \dot{y}_3 = 3 \cdot 10^7 y_2^2 \end{aligned}$$

We see that $\dot{y}_1 + \dot{y}_2 + \dot{y}_3 = 0$, hence the total mass $I(y) = y_1 + y_2 + y_3$ is an invariant of the system.

As was noted by Shampine (1986), such linear invariants are generally conserved by numerical integrators.

Theorem 1.5 (Conservation of Linear Invariants). *All explicit and implicit Runge–Kutta methods conserve linear invariants. Partitioned Runge–Kutta methods (II.2.2) conserve linear invariants if $b_i = \hat{b}_i$ for all i , or if the invariant depends only on p or only on q .*

Proof. Let $I(y) = d^T y$ with a constant vector d , so that $d^T f(y) = 0$ for all y . In the case of Runge–Kutta methods we thus have $d^T k_i = 0$, and consequently $d^T y_1 = d^T y_0 + h d^T (\sum_{i=1}^s b_i k_i) = d^T y_0$. The statement for partitioned methods is proved similarly. \square

Next we consider differential equations of the form

$$\dot{Y} = A(Y)Y, \quad (1.3)$$

where Y can be a vector or a matrix (not necessarily a square matrix). We then have the following result.

Theorem 1.6. *If $A(Y)$ is skew-symmetric for all Y (i.e., $A^T = -A$), then the quadratic function $I(Y) = Y^T Y$ is an invariant. In particular, if the initial value Y_0 consists of orthonormal columns (i.e., $Y_0^T Y_0 = I$), then the columns of the solution $Y(t)$ of (1.3) remain orthonormal for all t .*

Proof. The derivative of $I(Y)$ is $I'(Y)H = Y^T H + H^T Y$. Thus, we have $I'(Y)f(Y) = I'(Y)(A(Y)Y) = Y^T A(Y)Y + Y^T A(Y)^T Y$ for all Y which vanishes, because $A(Y)$ is skew-symmetric. This proves the statement. \square

Example 1.7 (Rigid Body). The motion of a free rigid body, whose centre of mass is at the origin, is described by the Euler equations

$$\begin{aligned} \dot{y}_1 &= a_1 y_2 y_3, & a_1 &= (I_2 - I_3)/(I_2 I_3) \\ \dot{y}_2 &= a_2 y_3 y_1, & a_2 &= (I_3 - I_1)/(I_3 I_1) \\ \dot{y}_3 &= a_3 y_1 y_2, & a_3 &= (I_1 - I_2)/(I_1 I_2) \end{aligned} \quad (1.4)$$

where the vector $y = (y_1, y_2, y_3)^T$ represents the angular momentum in the body frame, and I_1, I_2, I_3 are the principal moments of inertia (Euler (1758b); see Sect. VII.5 for a detailed description). This problem can be written as

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & y_3/I_3 & -y_2/I_2 \\ -y_3/I_3 & 0 & y_1/I_1 \\ y_2/I_2 & -y_1/I_1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, \quad (1.5)$$

which is of the form (1.3) with a skew-symmetric matrix $A(Y)$. By Theorem 1.6, $y_1^2 + y_2^2 + y_3^2$ is an invariant. A second quadratic invariant is

$$H(y_1, y_2, y_3) = \frac{1}{2} \left(\frac{y_1^2}{I_1} + \frac{y_2^2}{I_2} + \frac{y_3^2}{I_3} \right),$$

which represents the kinetic energy.

Inspired by the cover page of Marsden & Ratiu (1999), we present in Fig. 1.1 the sphere with some of the solutions of (1.4) corresponding to $I_1 = 2$, $I_2 = 1$ and $I_3 = 2/3$. They lie on the intersection of the sphere with the ellipsoid given by $H(y_1, y_2, y_3) = \text{Const.}$ In the left picture we have included the numerical solution (30 steps) obtained by the implicit midpoint rule with step size $h = 0.3$ and initial value $y_0 = (\cos(1.1), 0, \sin(1.1))^T$. It stays exactly on a solution curve. This follows from the fact that the implicit midpoint rule preserves quadratic invariants exactly (Sect. IV.2).

For the explicit Euler method (right picture of Fig. 1.1, 320 steps with $h = 0.05$ and the same initial value) we see that the numerical solution shows a wrong qualitative behaviour (it should lie on a closed curve). The numerical solution even drifts away from the sphere.

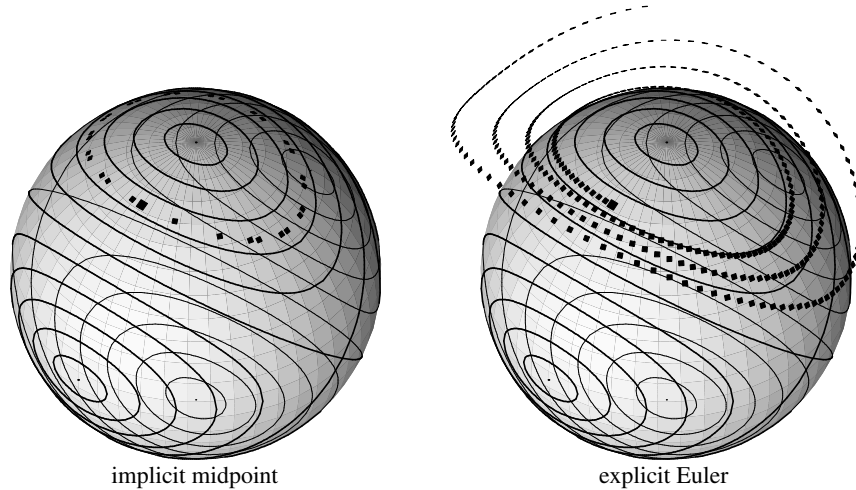


Fig. 1.1. Solutions of the Euler equations (1.4) for the rigid body

IV.2 Quadratic Invariants

Quadratic invariants appear often in applications. Examples are the conservation law of angular momentum in N -body systems (Example 1.3), the two invariants of the rigid body motion (Example 1.7), and the invariant $Y^T Y$ of Theorem 1.6. We therefore consider differential equations (1.1) and quadratic functions

$$Q(y) = y^T C y, \quad (2.1)$$

where C is a symmetric square matrix. It is an invariant of (1.1) if $y^T C f(y) = 0$ for all y .

IV.2.1 Runge–Kutta Methods

We shall give a complete characterization of Runge–Kutta methods which automatically conserve all quadratic invariants. We first of all consider the Gauss collocation methods.

Theorem 2.1. *The Gauss methods of Sect. II.1.3 (collocation based on the shifted Legendre polynomials) conserve quadratic invariants.*

Proof. Let $u(t)$ be the collocation polynomial of the Gauss methods (Definition II.1.3). Since $\frac{d}{dt}Q(u(t)) = 2u(t)^T C \dot{u}(t)$, it follows from $u(t_0) = y_0$ and $u(t_0 + h) = y_1$ that

$$y_1^T C y_1 - y_0^T C y_0 = 2 \int_{t_0}^{t_0+h} u(t)^T C \dot{u}(t) dt. \quad (2.2)$$

The integrand $u(t)^T C \dot{u}(t)$ is a polynomial of degree $2s - 1$, which is integrated without error by the s -stage Gaussian quadrature formula. It therefore follows from the collocation condition

$$u(t_0 + c_i h)^T C \dot{u}(t_0 + c_i h) = u(t_0 + c_i h)^T C f(u(t_0 + c_i h)) = 0$$

that the integral in (2.2) vanishes. \square

Since the implicit midpoint rule is the special case $s = 1$ of the Gauss methods, the preceding theorem explains its good behaviour for the rigid body simulation in Fig 1.1.

Theorem 2.2 (Cooper 1987). *If the coefficients of a Runge–Kutta method satisfy*

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \quad \text{for all } i, j = 1, \dots, s, \quad (2.3)$$

*then it conserves quadratic invariants.*²

² For irreducible methods, the conditions of Theorem 2.2 and Theorem 2.4 are also necessary for the conservation of all quadratic invariants. This follows from the discussion in Sect. VI.7.3.

Proof. The proof is the same as that for B-stability, given independently by Burrage & Butcher and Crouzeix in 1979 (see Hairer & Wanner (1996), Sect. IV.12).

The relation $y_1 = y_0 + h \sum_{i=1}^s b_i k_i$ of Definition II.1.1 yields

$$y_1^T C y_1 = y_0^T C y_0 + h \sum_{i=1}^s b_i k_i^T C y_0 + h \sum_{j=1}^s b_j y_0^T C k_j + h^2 \sum_{i,j=1}^s b_i b_j k_i^T C k_j. \quad (2.4)$$

We then write $k_i = f(Y_i)$ with $Y_i = y_0 + h \sum_{j=1}^s a_{ij} k_j$. The main idea is to compute y_0 from this relation and to insert it into the central expressions of (2.4). This yields (using the symmetry of C)

$$y_1^T C y_1 = y_0^T C y_0 + 2h \sum_{i=1}^s b_i Y_i^T C f(Y_i) + h^2 \sum_{i,j=1}^s (b_i b_j - b_i a_{ij} - b_j a_{ji}) k_i^T C k_j.$$

The condition (2.3) together with the assumption $y^T C f(y) = 0$, which states that $y^T C y$ is an invariant of (1.1), imply $y_1^T C y_1 = y_0^T C y_0$. \square

The criterion (2.3) is very restrictive. One finds that among all collocation and discontinuous collocation methods (Definition II.1.7) only the Gauss methods satisfy this criterion (Exercise 6). On the other hand, it is possible to construct other high-order Runge–Kutta methods satisfying (2.3). The key for such a construction is the W -transformation (see Hairer & Wanner (1996), Sect. IV.5), which is exploited in the articles of Sun (1993a) and Hairer & Leone (2000).

IV.2.2 Partitioned Runge–Kutta Methods

We next consider partitioned Runge–Kutta methods for systems $\dot{y} = f(y, z)$, $\dot{z} = g(y, z)$. Usually such methods cannot conserve general quadratic invariants (Exercise 4). We therefore concentrate on quadratic invariants of the form

$$Q(y, z) = y^T D z, \quad (2.5)$$

where D is a matrix of the appropriate dimensions. Observe that the angular momentum of N -body systems (Example 1.3) is of this form.

Theorem 2.3 (Sun 1993b). *The Lobatto IIIA - IIIB pair conserves all quadratic invariants of the form (2.5). In particular, this is true for the Störmer–Verlet scheme (see Sect. II.2.2).*

Proof. Let $u(t)$ and $v(t)$ be the (discontinuous) collocation polynomials of the Lobatto IIIA and Lobatto IIIB methods, respectively (see Sect. II.2.2). In analogy to the proof of Theorem 2.1 we have

$$\begin{aligned} & Q(u(t_0 + h), v(t_0 + h)) - Q(u(t_0), v(t_0)) \\ &= \int_{t_0}^{t_0+h} \left(Q(\dot{u}(t), v(t)) + Q(u(t), \dot{v}(t)) \right) dt. \end{aligned} \quad (2.6)$$

Since $u(t)$ is of degree s and $v(t)$ of degree $s - 2$, the integrand of (2.6) is a polynomial of degree $2s - 3$. Hence, an application of the Lobatto quadrature yields the exact result. Using the fact that $Q(y, z)$ is an invariant of the differential equation, i.e., $Q(f(y, z), z) + Q(y, g(y, z)) \equiv 0$, we thus obtain for the integral in (2.6)

$$hb_1 Q(u(t_0), \delta(t_0)) + hb_s Q(u(t_0 + h), \delta(t_0 + h)),$$

where $\delta(t) = \dot{v}(t) - g(u(t), v(t))$ denotes the defect. It now follows from $u(t_0) = y_0$, $u(t_0 + h) = y_1$ (definition of Lobatto IIIA) and from $v(t_0) = z_0 - hb_1\delta(t_0)$, $v(t_0 + h) = z_1 + hb_s\delta(t_0 + h)$ (definition of Lobatto IIIB) that $Q(y_1, z_1) - Q(y_0, z_0) = 0$, which proves the theorem. \square

Exchanging the role of the IIIA and IIIB methods also leads to an integrator that preserves quadratic invariants of the form (2.5). The following characterization extends Theorem 2.2 to partitioned Runge–Kutta methods.

Theorem 2.4. *If the coefficients of a partitioned Runge–Kutta method (II.2.2) satisfy*

$$b_i \hat{a}_{ij} + \hat{b}_j a_{ji} = b_i \hat{b}_j \quad \text{for } i, j = 1, \dots, s, \quad (2.7)$$

$$b_i = \hat{b}_i \quad \text{for } i = 1, \dots, s, \quad (2.8)$$

then it conserves quadratic invariants of the form (2.5).

If the partitioned differential equation is of the special form $\dot{y} = f(z)$, $\dot{z} = g(y)$, then condition (2.7) alone implies that invariants of the form (2.5) are conserved.

Proof. The proof is nearly identical to that of Theorem 2.2. Instead of (2.4) we get

$$y_1^T D z_1 = y_0^T D z_0 + h \sum_{i=1}^s b_i k_i^T D z_0 + h \sum_{j=1}^s \hat{b}_j y_0^T D \ell_j + h^2 \sum_{i,j=1}^s b_i \hat{b}_j k_i^T D \ell_j.$$

Denoting by (Y_i, Z_i) the arguments of $k_i = f(Y_i, Z_i)$ and $\ell_i = g(Y_i, Z_i)$, the same trick as in the proof of Theorem 2.2 gives

$$\begin{aligned} y_1^T D z_1 &= y_0^T D z_0 + h \sum_{i=1}^s b_i f(Y_i, Z_i)^T D Z_i + h \sum_{j=1}^s \hat{b}_j Y_j^T D g(Y_j, Z_j) \\ &\quad + h^2 \sum_{i,j=1}^s (b_i \hat{b}_j - b_i \hat{a}_{ij} - \hat{b}_j a_{ji}) k_i^T D \ell_j. \end{aligned} \quad (2.9)$$

Since (2.5) is an invariant, we have $f(y, z)^T D z + y^T D g(y, z) = 0$ for all y and z . Consequently, the two conditions (2.7) and (2.8) imply $y_1^T D z_1 = y_0^T D z_0$.

For the special case where f depends only on z and g only on y , the assumption $f(z)^T D z + y^T D g(y) = 0$ (for all y, z) implies that $f(z)^T D z = -y^T D g(y) = \text{Const.}$ Therefore, condition (2.8) is no longer necessary for the proof of the statement. \square

IV.2.3 Nyström Methods

An important class of partitioned differential equations is $\dot{y} = z$, $\dot{z} = g(y)$ or, equivalently,

$$\ddot{y} = g(y). \quad (2.10)$$

Many examples of Chap. I are of this form, in particular the N -body problem of Example 1.3 for which the angular momentum is a quadratic first integral. Nyström methods (Definition II.2.3),

$$\begin{aligned} \ell_i &= g\left(y_0 + c_i h \dot{y}_0 + h^2 \sum_{j=1}^s a_{ij} \ell_j\right), \\ y_1 &= y_0 + h \dot{y}_0 + h^2 \sum_{i=1}^s \beta_i \ell_i, \quad \dot{y}_1 = \dot{y}_0 + h \sum_{i=1}^s b_i \ell_i, \end{aligned} \quad (2.11)$$

are adapted to the numerical solution of (2.10) and it is interesting to investigate which methods within this class can conserve quadratic invariants.

Theorem 2.5. *If the coefficients of the Nyström method (2.11) satisfy*

$$\begin{aligned} \beta_i &= b_i(1 - c_i) \quad \text{for } i = 1, \dots, s, \\ b_i(\beta_j - a_{ij}) &= b_j(\beta_i - a_{ji}) \quad \text{for } i, j = 1, \dots, s, \end{aligned} \quad (2.12)$$

then it conserves all quadratic invariants of the form $y^T D \dot{y}$.

Proof. The quadratic form $Q(y, \dot{y}) = y^T D \dot{y}$ is a first integral of (2.10) if and only if

$$\dot{y}^T D \dot{y} + y^T D g(y) = 0 \quad \text{for all } y, \dot{y} \in \mathbb{R}^n. \quad (2.13)$$

This implies that D is skew-symmetric and that $y^T D g(y) = 0$.

In the same way as for the proofs of Theorems 2.2 and 2.4 we now compute $y_1^T D \dot{y}_1$ using the formulas of (2.11) and we substitute y_0 by $Y_i - c_i h \dot{y}_0 - h^2 \sum_j a_{ij} \ell_j$, where Y_i denotes the argument of g in (2.11). This yields

$$\begin{aligned} y_1^T D \dot{y}_1 &= y_0^T D \dot{y}_0 + h \dot{y}_0^T D \dot{y}_0 + h \sum_{i=1}^s b_i Y_i^T D \ell_i \\ &+ h^2 \sum_{i=1}^s \beta_i \ell_i^T D \dot{y}_0 + h^2 \sum_{i=1}^s b_i(1 - c_i) \dot{y}_0^T D \ell_i \\ &+ h^3 \sum_{i,j=1}^s b_i(\beta_j - a_{ij}) \ell_j^T D \ell_i. \end{aligned}$$

Using the skew-symmetry of D and $Y_i^T D \ell_i = Y_i^T D g(Y_i) = 0$, condition (2.12) implies the conservation property $y_1^T D \dot{y}_1 = y_0^T D \dot{y}_0$. \square

Remark 2.6 (Composition Methods). If a method Φ_h conserves quadratic invariants (e.g., the mid-point rule by Theorem 2.1 or the Störmer–Verlet scheme by Theorem 2.3 or a Nyström method of Theorem 2.5), then so does the composition method

$$\Psi_h = \Phi_{\gamma_s h} \circ \dots \circ \Phi_{\gamma_1 h}. \quad (2.14)$$

This obvious property is one of the most important motivations for considering composition methods.

IV.3 Polynomial Invariants

We consider two classes of problems with polynomial invariants for degree higher than two. First, we treat linear problems for which the determinant of the resolvent is an invariant, and we show that (partitioned) Runge–Kutta methods cannot conserve them automatically. Second, we study isospectral flows.

IV.3.1 The Determinant as a First Integral

We consider quasi-linear problems

$$\dot{Y} = A(Y)Y, \quad Y(0) = Y_0 \quad (3.1)$$

where Y and $A(Y)$ are $n \times n$ matrices. In the following we denote the trace of a matrix $A = (a_{ij})_{i,j=1}^n$ by $\text{trace } A = \sum_{i=1}^n a_{ii}$.

Lemma 3.1. *If $\text{trace } A(Y) = 0$ for all Y , then $g(Y) := \det Y$ is an invariant of the matrix differential equation (3.1).*

Proof. It follows from

$$\det(Y + \varepsilon AY) = \det(I + \varepsilon A) \det Y = (1 + \varepsilon \text{trace } A + \mathcal{O}(\varepsilon^2)) \det Y$$

that $g'(Y)(AY) = \text{trace } A \cdot \det Y$ (this is the *Abel–Liouville–Jacobi–Ostrogradskii identity*). Hence, the determinant $g(Y) = \det Y$ is an invariant of the differential equation (3.1) if $\text{trace } A(Y) = 0$ for all Y . \square

Since $\det Y$ represents the volume of the parallelepiped generated by the columns of the matrix Y , the conservation of the invariant $g(Y) = \det Y$ is related to volume preservation. This topic will be further discussed in Sect. VI.9. Here, we consider $\det Y$ as a polynomial invariant of degree n , and we investigate whether Runge–Kutta methods can automatically conserve this invariant for $n \geq 3$. The key lemma for this study is the following.

Lemma 3.2 (Feng Kang & Shang Zai-jiu 1995). *Let $R(z)$ be a differentiable function defined in a neighbourhood of $z = 0$, and assume that $R(0) = 1$ and $R'(0) = 1$. Then, we have for $n \geq 3$*

$$\det R(A) = 1 \quad \text{for all } n \times n \text{ matrices } A \text{ satisfying } \text{trace } A = 0, \quad (3.2)$$

if and only if $R(z) = \exp(z)$.

Proof. The “if” part follows from Lemma 3.1, because for constant A the solution of $\dot{Y} = AY$, $Y(0) = I$ is given by $Y(t) = \exp(At)$.

For the proof of the “only if” part, we consider diagonal matrices of the form $A = \text{diag}(\mu, \nu, -(\mu + \nu), 0, \dots, 0)$, which have $\text{trace } A = 0$, and for which

$$R(A) = \text{diag}(R(\mu), R(\nu), R(-(\mu + \nu)), R(0), \dots, R(0)).$$

The assumptions $R(0) = 1$ and (3.2) imply

$$R(\mu)R(\nu)R(-(\mu + \nu)) = 1 \quad (3.3)$$

for all μ, ν close to 0. Putting $\nu = 0$, this relation yields $R(\mu)R(-\mu) = 1$ for all μ , and therefore (3.3) can be written as

$$R(\mu)R(\nu) = R(\mu + \nu) \quad \text{for all } \mu, \nu \text{ close to } 0. \quad (3.4)$$

This functional equation can only be satisfied by the exponential function. This is seen as follows: from (3.4) we have

$$\frac{R(\mu + \varepsilon) - R(\mu)}{\varepsilon} = R(\mu) \frac{R(\varepsilon) - R(0)}{\varepsilon}.$$

Taking the limit $\varepsilon \rightarrow 0$ we obtain $R'(\mu) = R(\mu)$, because $R'(0) = 1$. This implies $R(\mu) = \exp(\mu)$. \square

Theorem 3.3. *For $n \geq 3$, no Runge–Kutta method can conserve all polynomial invariants of degree n .*

Proof. It is sufficient to consider linear problems $\dot{Y} = AY$ with constant matrix A satisfying $\text{trace } A = 0$, so that $g(Y) = \det Y$ is a polynomial invariant of degree n . Applying a Runge–Kutta method to such a differential equation yields $Y_1 = R(hA)Y_0$, where

$$R(z) = 1 + zb^T(I - z\mathcal{A})^{-1}\mathbb{1}$$

($b^T = (b_1, \dots, b_s)$, $\mathbb{1} = (1, \dots, 1)^T$ and $\mathcal{A} = (a_{ij})$ is the matrix of Runge–Kutta coefficients) is the so-called stability function. It is seen to be rational. By Lemma 3.2 it is therefore not possible that $\det R(hA) = 1$ for all A with $\text{trace } A = 0$. \square

This negative result motivates the search for new methods which can conserve polynomial invariants (see Sects. IV.4, IV.8 and VI.9). We consider here another interesting class of problems with polynomial invariants of degree higher than two.

IV.3.2 Isospectral Flows

Such flows are created by a matrix differential equation

$$\dot{L} = [B(L), L], \quad L(0) = L_0 \quad (3.5)$$

where L_0 is a given symmetric matrix, $B(L)$ is skew-symmetric for all L , and $[B, L] = BL - LB$ is the commutator of B and L . Many interesting problems can be written in this form. We just mention the Toda system, the continuous realization of QR-type algorithms, projected gradient flows, and inverse eigenvalue problems (see Chu (1992) and Calvo, Iserles & Zanna (1997) for long lists of references).

Lemma 3.4 (Lax 1968, Flaschka 1974). *Let L_0 be symmetric and assume that $B(L)$ is skew-symmetric for all L . Then, the solution $L(t)$ of (3.5) is a symmetric matrix, and its eigenvalues are independent of t .*

Proof. The symmetry of $L(t)$ follows from the fact that the commutator of a skew-symmetric with a symmetric matrix gives a symmetric matrix.

To prove the isospectrality of the flow, we define $U(t)$ by

$$\dot{U} = B(L(t)) U, \quad U(0) = I. \quad (3.6)$$

Then, we have $(d/dt)(U^{-1}LU) = U^{-1}(\dot{L} - BL + LB)U = 0$, and hence $U(t)^{-1}L(t)U(t) = L_0$ for all t , so that $L(t) = U(t)L_0U(t)^{-1}$ is the solution of (3.5). This proves the result. \square

Note that, since $B(L)$ is skew-symmetric, the matrix $U(t)$ of (3.6) is orthogonal by Theorem 1.6.

Lemma 3.4 shows that the characteristic polynomial $\det(L - \lambda I) = \sum_{i=0}^n a_i \lambda^i$ and hence the coefficients a_i also are independent of t . These coefficients are all polynomial invariants (e.g., $a_0 = \det L$, $a_{n-1} = \pm \text{trace } L$). Because of Theorem 3.3 there is no hope that Runge–Kutta methods applied to (3.5) can conserve these invariants automatically for $n \geq 3$.

Isospectral Methods. The proof of Lemma 3.4, however, suggests an interesting approach for the numerical solution of (3.5). For $n = 0, 1, \dots$ we solve numerically

$$\dot{U} = B(UL_nU^T)U, \quad U(0) = I \quad (3.7)$$

and we put $L_{n+1} = \hat{U}L_n\hat{U}^T$, where \hat{U} is the numerical approximation $\hat{U} \approx U(h)$ after one step (cf. Calvo, Iserles & Zanna 1999). If $B(L)$ is skew-symmetric for all matrices L , then U^TU is a quadratic invariant of (3.7) and the methods of Sect. IV.2 will produce an orthogonal \hat{U} . Consequently, L_{n+1} and L_n have exactly the same eigenvalues, and they remain symmetric.

Diele, Lopez & Politi (1998) suggest the use of the Cayley transform $U = (I - Y)^{-1}(I + Y)$, which transforms (3.7) into

$$\dot{Y} = \frac{1}{2}(I - Y)B(UL_nU^T)(I + Y), \quad Y(0) = 0,$$

and the orthogonality of U into the skew-symmetry of Y (see Lemma 8.8 below). Since all (also explicit) Runge–Kutta methods preserve the skew-symmetry of Y , which is a linear invariant, this yields an approach to explicit isospectral methods.

Connection with the QR Algorithm. In a diversion from the main theme of this section, we now show the relationship of the flow of (3.5) with the QR algorithm for the symmetric eigenvalue problem. Starting from a real symmetric matrix A_0 , the basic *QR algorithm* (without shifts) computes a sequence of orthogonally similar matrices A_1, A_2, A_3, \dots , expected to converge towards a diagonal matrix carrying the eigenvalues of A_0 . Iteratively for $k = 0, 1, 2, \dots$, one computes the QR decomposition of A_k :

$$A_k = Q_k R_k$$

with Q_k orthogonal, R_k upper triangular (the decomposition becomes unique if the diagonal elements of R_k are taken positive). Then, A_{k+1} is obtained by reversing the order of multiplication:

$$A_{k+1} = R_k Q_k.$$

It is an easy exercise to show that $Q(k) = Q_0 Q_1 \dots Q_{k-1}$ is the matrix in the orthogonal similarity transformation between A_0 and A_k :

$$A_k = Q(k)^T A_0 Q(k) \quad (3.8)$$

and the same matrix $Q(k)$ is the orthogonal factor in the QR decomposition of A_0^k :

$$A_0^k = Q(k) R(k). \quad (3.9)$$

Consider now, for an arbitrary real function f defined on the eigenvalues of a real symmetric matrix L_0 , the QR decomposition

$$\exp(tf(L_0)) = Q(t) R(t) \quad (3.10)$$

and define

$$L(t) := Q(t)^T L_0 Q(t). \quad (3.11)$$

The relations (3.8) and (3.9) then show that for integer times $t = k$, the matrix $\exp(f(L(k))) = Q(k)^T \exp(f(L_0)) Q(k)$ coincides with the k th matrix in the QR algorithm starting from $A_0 = \exp(f(L_0))$:

$$\exp(f(L(k))) = A_k. \quad (3.12)$$

Now, how is all this related to the system (3.5)? Differentiating (3.11) as in the proof of Lemma 3.4 shows that $L(t)$ solves a differential equation of the form $\dot{L} = [B, L]$ with the skew-symmetric matrix $B = -Q^T \dot{Q}$. At first sight, however, B is a function of t , not of L . On the other hand, differentiation of (3.10) yields (omitting the argument t where it is clear from the context)

$$f(L_0)QR = f(L_0) \exp(tf(L_0)) = \exp(tf(L_0))f(L_0) = \dot{Q}R + Q\dot{R},$$

and since $f(L) = Q^T f(L_0) Q$ by (3.11), this becomes

$$f(L) = Q^T \dot{Q} + \dot{R} R^{-1}.$$

Here the left-hand side is a symmetric matrix, and the right-hand side is the sum of a skew-symmetric and an upper triangular matrix. It follows that the skew-symmetric matrix $B = -Q^T \dot{Q}$ is given by

$$B(L) = f(L)_+ - f(L)_+^T, \quad (3.13)$$

where $f(L)_+$ denotes the part of $f(L)$ above the diagonal. Hence, $L(t)$ is the solution of an autonomous system (3.5) with a skew-symmetric $B(L)$.

For $f(x) = x$ and assuming L_0 symmetric and tridiagonal, the flow of (3.5) with (3.13) is known as the *Toda flow*. The QR iterates $A_0 = \exp(L_0)$, A_1, A_2, \dots of the exponential of L_0 are seen to be equal to the exponentials of the solution $L(t)$ of the Toda equations at integer times: $A_k = \exp(L(k))$, a discovery of Symes (1982). An interesting connection of the Toda equations with a mechanical system will be discussed in Sect. X.1.5.

For $f(x) = \log x$, the above arguments show that the QR iteration itself, starting from a positive definite symmetric tridiagonal matrix, is the evaluation $A_k = L(k)$ at integer times of a solution $L(t)$ of the differential equation (3.5) with B given by (3.13). This relationship was explored in a series of papers by Deift, Li, Nanda & Tomei (1983, 1989, 1993).

Notwithstanding the mathematical beauty of this relationship, it must be remarked that the practical QR algorithm (with shifts and deflation) follows a different path.

IV.4 Projection Methods

Und bist du nicht willig, so brauch ich Gewalt.

(J.W. Goethe, *Der Erlkönig*)

Suppose we have an $(n - m)$ -dimensional submanifold of \mathbb{R}^n ,

$$\mathcal{M} = \{y ; g(y) = 0\} \quad (4.1)$$

($g : \mathbb{R}^n \rightarrow \mathbb{R}^m$), and a differential equation $\dot{y} = f(y)$ with the property that

$$y_0 \in \mathcal{M} \quad \text{implies} \quad y(t) \in \mathcal{M} \quad \text{for all } t. \quad (4.2)$$

We want to emphasize that this assumption is weaker than the requirement that all components $g_i(y)$ of $g(y)$ are invariants in the sense of Definition 1.1. In fact, assumption (4.2) is equivalent to $g'(y)f(y) = 0$ for $y \in \mathcal{M}$, whereas Definition 1.1 requires $g'(y)f(y) = 0$ for all $y \in \mathbb{R}^n$. In the situation of (4.2) we call $g(y)$ a *weak invariant*, and we say that $\dot{y} = f(y)$ is a differential equation on the manifold \mathcal{M} .

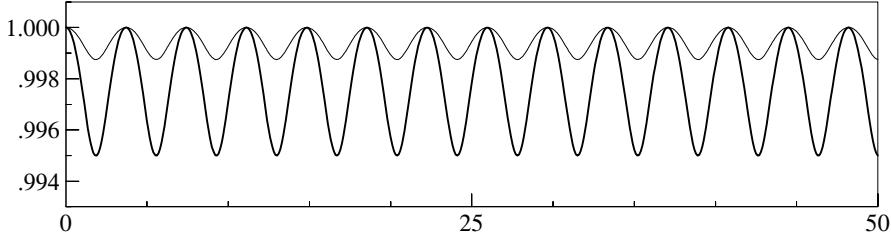


Fig. 4.1. The implicit midpoint rule applied to the differential equation (4.3). The picture shows the numerical values for $q_1^2 + q_2^2$ obtained with step size $h = 0.1$ (thick line) and $h = 0.05$ (thin line)

Example 4.1. Consider the pendulum equation written in Cartesian coordinates:

$$\begin{aligned} \dot{q}_1 &= p_1, & \dot{p}_1 &= -q_1 \lambda, \\ \dot{q}_2 &= p_2, & \dot{p}_2 &= -1 - q_2 \lambda, \end{aligned} \quad (4.3)$$

where $\lambda = (p_1^2 + p_2^2 - q_2)/(q_1^2 + q_2^2)$. One can check by differentiation that $q_1 p_1 + q_2 p_2$ (orthogonality of the position and velocity vectors) is an invariant in the sense of Definition 1.1. However, $q_1^2 + q_2^2$ (length of the pendulum) is only a weak invariant. The experiment of Fig. 4.1 shows that even methods which conserve quadratic first integrals (cf. Sect. IV.2) do not conserve the quadratic weak invariant $q_1^2 + q_2^2$. No numerical method that is allowed to evaluate the vector field $f(y)$ outside \mathcal{M} can be expected to conserve weak invariants exactly. This is one of the motivations for considering the methods of this and the subsequent sections.

A natural approach to the numerical solution of differential equations on manifolds is by projection (see e.g., Hairer & Wanner (1996), Sect. VII.2, Eich-Soellner & Führer (1998), Sect. 5.3.3).

Algorithm 4.2 (Standard Projection Method). Assume that $y_n \in \mathcal{M}$. One step $y_n \mapsto y_{n+1}$ is defined as follows (see Fig. 4.2):

- Compute $\tilde{y}_{n+1} = \Phi_h(y_n)$, where Φ_h is an arbitrary one-step method applied to $\dot{y} = f(y)$;
- project the value \tilde{y}_{n+1} onto the manifold \mathcal{M} to obtain $y_{n+1} \in \mathcal{M}$.

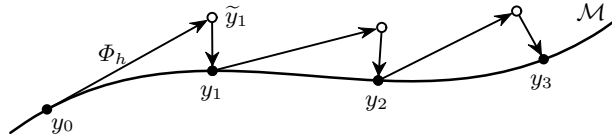


Fig. 4.2. Illustration of the standard projection method

For $y_n \in \mathcal{M}$ the distance of \tilde{y}_{n+1} to the manifold \mathcal{M} is of the size of the local error, i.e., $\mathcal{O}(h^{p+1})$. Therefore, the projection does not deteriorate the convergence order of the method.

For the computation of y_{n+1} we have to solve the constrained minimization problem

$$\|y_{n+1} - \tilde{y}_{n+1}\| \rightarrow \min \quad \text{subject to} \quad g(y_{n+1}) = 0. \quad (4.4)$$

In the case of the Euclidean norm, a standard approach is to introduce Lagrange multipliers $\lambda = (\lambda_1, \dots, \lambda_m)^T$, and to consider the Lagrange function $\mathcal{L}(y_{n+1}, \lambda) = \|y_{n+1} - \tilde{y}_{n+1}\|^2/2 - g(y_{n+1})^T \lambda$. The necessary condition $\partial \mathcal{L} / \partial y_{n+1} = 0$ then leads to the system

$$\begin{aligned} y_{n+1} &= \tilde{y}_{n+1} + g'(\tilde{y}_{n+1})^T \lambda \\ 0 &= g(y_{n+1}). \end{aligned} \quad (4.5)$$

We have replaced y_{n+1} with \tilde{y}_{n+1} in the argument of $g'(y)$ in order to save some evaluations of $g'(y)$. Inserting the first relation of (4.5) into the second gives a non-linear equation for λ , which can be efficiently solved by simplified Newton iterations:

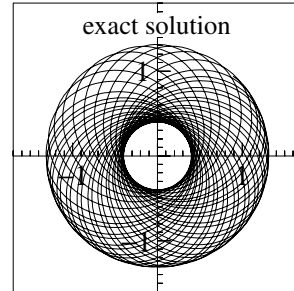
$$\Delta \lambda_i = - \left(g'(\tilde{y}_{n+1}) g'(\tilde{y}_{n+1})^T \right)^{-1} g \left(\tilde{y}_{n+1} + g'(\tilde{y}_{n+1})^T \lambda_i \right), \quad \lambda_{i+1} = \lambda_i + \Delta \lambda_i.$$

For the choice $\lambda_0 = 0$ the first increment $\Delta \lambda_0$ is of size $\mathcal{O}(h^{p+1})$, so that the convergence is usually extremely fast. Often, one simplified Newton iteration is sufficient.

Example 4.3. As a first example we consider the perturbed Kepler problem (see Exercise I.12) with Hamiltonian function

$$\begin{aligned} H(p, q) &= \frac{1}{2} (p_1^2 + p_2^2) - \frac{1}{\sqrt{q_1^2 + q_2^2}} \\ &\quad - \frac{0.005}{2\sqrt{(q_1^2 + q_2^2)^3}}, \end{aligned}$$

and initial values $q_1(0) = 1 - e$, $q_2(0) = 0$, $p_1(0) = 0$, $p_2(0) = \sqrt{(1+e)/(1-e)}$ (eccentricity $e = 0.6$) on the interval $0 \leq t \leq 200$. The exact



solution (plotted to the right) is approximately an ellipse that rotates slowly around one of its foci. For this problem we know two first integrals: the Hamiltonian function $H(p, q)$ and the angular momentum $L(p, q) = q_1 p_2 - q_2 p_1$.

We apply the explicit Euler method and the symplectic Euler method (I.1.9), both with constant step size $h = 0.03$. The result is shown in Fig. 4.3. The numerical solution of the explicit Euler method (without projection) is completely wrong. The projection onto the manifold $\{H(p, q) = H(p_0, q_0)\}$ improves the numerical solution, but it still has a wrong qualitative behaviour. Only projection onto both invariants, $H(p, q) = \text{Const}$ and $L(p, q) = \text{Const}$ gives the correct behaviour. The symplectic Euler method already shows the correct behaviour without any projections (see Chap. IX for an explanation). Surprisingly, a projection onto $H(p, q) = \text{Const}$ destroys this behaviour, the numerical solution approaches the centre and the simplified Newton iterations fail to converge beyond $t = 25.23$. Projection onto both invariants re-establishes the correct behaviour.

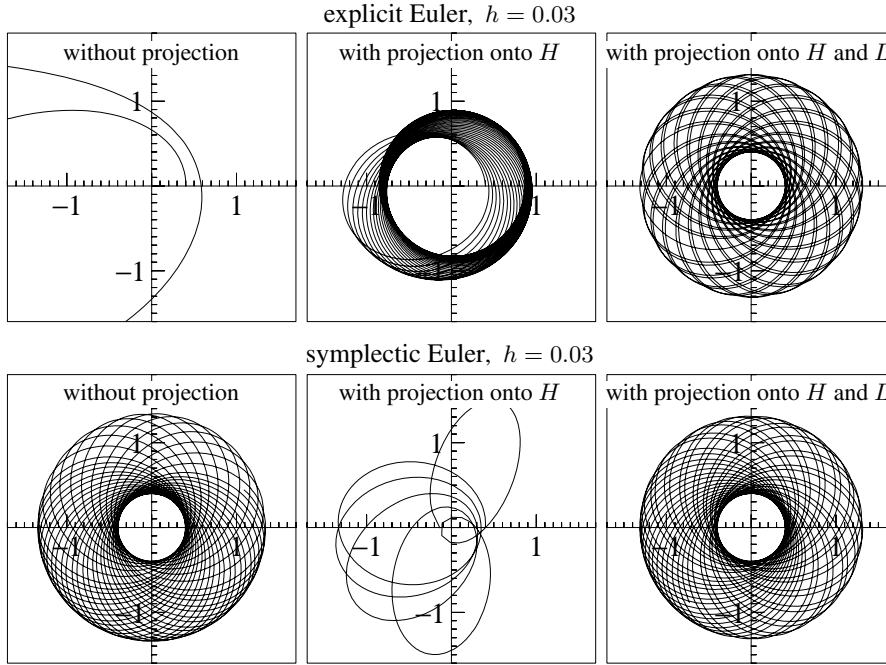


Fig. 4.3. Numerical solutions obtained with and without projections

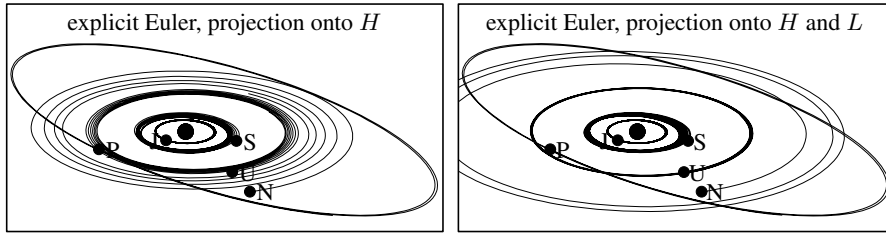


Fig. 4.4. Explicit Euler method with projections applied to the outer solar system, step size $h = 10$ (days), interval $0 \leq t \leq 200\,000$

Example 4.4 (Outer Solar System). Having encountered excellent experience with projections onto H and L for the perturbed Kepler problem (Example 4.3), let us apply the same idea to a more realistic problem in celestial mechanics. We consider the outer solar system as described in Sect. I.2. The numerical solution of the explicit Euler method applied with constant step size $h = 10$, once with projection onto $H = \text{Const}$ and once with projection onto $H = \text{Const}$ and $L = \text{Const}$, is shown in Fig. 4.4 (observe that the conservation of the angular momentum $L(p, q) = \sum_{i=1}^N q_i \times p_i$ consists of three first integrals). We see a slight improvement in the orbits of Jupiter, Saturn and Uranus (compared to the explicit

Euler method without projections, see Fig. I.2.4), but the orbit of Neptune becomes even worse. There is no doubt that this problem contains a structure which cannot be correctly simulated by methods that only preserve the total energy H and the angular momentum L .

Example 4.5 (Volume Preservation). Consider the matrix differential equation $\dot{Y} = A(Y)Y$, where $\text{trace } A(Y) = 0$ for all Y . We know from Lemma 3.1 that $g(Y) = \det Y$ is an invariant which cannot be automatically conserved by Runge–Kutta methods. Here, we show how we can enforce this invariant by projection. Let \tilde{Y}_{n+1} be the numerical approximation obtained with an arbitrary one-step method. We consider the Frobenius norm $\|Y\|_F = \sqrt{\sum_{i,j} |y_{ij}|^2}$ for measuring the distance to the manifold $\{Y; g(Y) = 0\}$. Using $g'(Y)(AY) = \text{trace } A \det Y$ (see the proof of Lemma 3.1) with A chosen such that the product AY contains only one non-zero element, the projection step (4.5) is seen to become (Exercise 9)

$$Y_{n+1} = \tilde{Y}_{n+1} + \mu \tilde{Y}_{n+1}^{-T} \quad (4.6)$$

with the scalar $\mu = \lambda \det \tilde{Y}_{n+1}$. This leads to the scalar nonlinear equation $\det(\tilde{Y}_{n+1} + \mu \tilde{Y}_{n+1}^{-T}) = \det Y_n$, for which simplified Newton iterations become

$$\det(\tilde{Y}_{n+1} + \mu_i \tilde{Y}_{n+1}^{-T}) \left(1 + (\mu_{i+1} - \mu_i) \text{trace}((\tilde{Y}_{n+1}^T \tilde{Y}_{n+1})^{-1})\right) = \det Y_n.$$

If the QR -decomposition of \tilde{Y}_{n+1} is available from the computation of $\det \tilde{Y}_{n+1}$, the value of $\text{trace}((\tilde{Y}_{n+1}^T \tilde{Y}_{n+1})^{-1})$ can be computed efficiently with $\mathcal{O}(n^3/3)$ flops (see e.g., Golub & Van Loan (1989), Sect. 5.3.9).

The above projection is preferable to $Y_{n+1} = c \tilde{Y}_{n+1}$, where $c \in \mathbb{R}$ is chosen such that $\det Y_{n+1} = \det Y_n$. This latter projection is already ill-conditioned for diagonal matrices with entries that differ by several magnitudes.

As a conclusion to the above numerical experiments we see that a projection can give excellent results, but can also destroy the good long-time behaviour of the solution if applied inappropriately. If the original method already preserves some structure, then projection to a subset of invariants may destroy the good long-time behaviour. An important modification for reversible differential equations (symmetric projections) will be presented in Sect. V.4.1.

IV.5 Numerical Methods Based on Local Coordinates

A second important class of methods for the numerical treatment of differential equations on manifolds uses local coordinates. Before explaining the ideas, we find it appropriate to discuss in more detail manifolds and differential equations on manifolds.

IV.5.1 Manifolds and the Tangent Space

In Sect. IV.4 we assumed that locally (in a neighbourhood U of $a \in \mathbb{R}^n$) a manifold is given by constraints, i.e.,

$$\mathcal{M} = \{y \in U ; g(y) = 0\}, \quad (5.1)$$

where $g : U \rightarrow \mathbb{R}^m$ is differentiable, $g(a) = 0$, and $g'(a)$ has full rank m .

Here, we use local parameters to characterize a manifold. Let $\psi : V \rightarrow \mathbb{R}^n$ be differentiable ($V \subset \mathbb{R}^{n-m}$ is a neighbourhood of 0), $\psi(0) = a$, and assume that $\psi'(0)$ has full rank $n - m$. Then, a manifold is locally given by

$$\mathcal{M} = \{y = \psi(z) ; z \in V\} \quad (5.2)$$

provided that V is sufficiently small, so that $\psi : V \rightarrow \psi(V)$ is bijective with continuous inverse. The variables z are called *parameters* or *local coordinates* of the manifold.

As an example, consider the unit sphere which, in the form (5.1), is given by the function $g(y_1, y_2, y_3) = y_1^2 + y_2^2 + y_3^2 - 1$. There are many possible choices of local coordinates. Away from the equator (i.e., $y_3 = 0$), we can take $z = (z_1, z_2)^T := (y_1, y_2)^T$ and $\psi(z) = (z_1, z_2, \pm\sqrt{1 - z_1^2 - z_2^2})^T$. Alternatively, we can consider spherical coordinates $\psi(\alpha, \beta) = (\cos \alpha \sin \beta, \sin \alpha \sin \beta, \cos \beta)^T$ away from the north and south poles (i.e., $y_1 = y_2 = 0, y_3 = \pm 1$).

The tangent to a curve (or the tangent plane to a surface) is an affine space passing through the contact point $a \in \mathcal{M}$. It is convenient to place the origin at a , so that we obtain a vector space. More precisely, for a manifold \mathcal{M} we define the *tangent space* at $a \in \mathcal{M}$ as

$$T_a \mathcal{M} = \left\{ v \in \mathbb{R}^n \mid \begin{array}{l} \text{there exists a differentiable path } \gamma : (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^n \\ \text{with } \gamma(t) \in \mathcal{M} \text{ for all } t, \gamma(0) = a, \dot{\gamma}(0) = v \end{array} \right\}. \quad (5.3)$$

Lemma 5.1. *If the manifold \mathcal{M} is given by (5.1), where $g : U \rightarrow \mathbb{R}^m$ is differentiable, $g(a) = 0$, and $g'(a)$ has full rank m , then we have*

$$T_a \mathcal{M} = \ker g'(a) = \{v \in \mathbb{R}^n \mid g'(a)v = 0\}. \quad (5.4)$$

If \mathcal{M} is given by (5.2), where $\psi : V \rightarrow \mathbb{R}^n$ is differentiable, $\psi(0) = a$, and $\psi'(0)$ has full rank $n - m$, then we have

$$T_a \mathcal{M} = \text{Im } \psi'(0) = \{\psi'(0)w \mid w \in \mathbb{R}^{n-m}\}. \quad (5.5)$$

Proof. a) For a path $\gamma(t)$ satisfying $\gamma(0) = a$ and $g(\gamma(t)) = 0$ it follows by differentiation that $g'(a)\dot{\gamma}(0) = 0$. Consequently, we have $T_a \mathcal{M} \subset \ker g'(a)$.

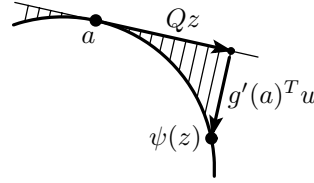
Consider now the function $F(t, u) = g(a + tv + g'(a)^T u)$. We have $F(0, 0) = 0$ and an invertible $\partial F / \partial u(0, 0) = g'(a)g'(a)^T$, so that by the implicit function theorem the relation $F(t, u) = 0$ can be solved locally for $u = u(t)$. If $v \in \ker g'(a)$,

it follows that $\dot{u}(0) = 0$, and the path $\gamma(t) = a + tv + g'(a)^T u(t)$ satisfies all requirements of (5.3), so that also $T_a \mathcal{M} \supset \ker g'(a)$.

b) Assume \mathcal{M} to be given by (5.2). For an arbitrary $\eta : (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^n$ satisfying $\eta(0) = 0$, the path $\gamma(t) = \psi(\eta(t))$ lies in \mathcal{M} and satisfies $\dot{\gamma}(0) = \psi'(0)\dot{\eta}(0)$. This proves $\text{Im } \psi'(0) \subset T_a \mathcal{M}$.

The assumption on the rank of $\psi'(0)$ implies that, after a reordering of the components, we have $\psi(z) = (\psi_1(z), \psi_2(z))^T$, where $\psi_1(z)$ is a local diffeomorphism (by the inverse function theorem). We show that every smooth path $\gamma(t)$ in \mathcal{M} can be written as $\gamma(t) = \psi(\eta(t))$ with some smooth $\eta(t)$. This then implies $T_a \mathcal{M} \subset \text{Im } \psi'(0)$. To prove this we split $\gamma(t) = (\gamma_1(t), \gamma_2(t))^T$ according to the partitioning of ψ , and we define $\eta(t) = \psi_1^{-1}(\gamma_1(t))$. Since for $\gamma(t) \in \mathcal{M}$ the second part $\gamma_2(t)$ is uniquely determined by $\gamma_1(t)$, this proves $\gamma(t) = \psi(\eta(t))$. \square

The proof of the preceding lemma shows the equivalence of the representations (5.1) and (5.2) of manifolds in \mathbb{R}^n . Let \mathcal{M} be given by (5.1), and assume that the columns of Q form an orthogonal basis of $T_a \mathcal{M}$. As in part (a) of the proof of Lemma 5.1 the condition $g(a + Qz + g'(a)^T u) = 0$ defines locally (close to $z = 0$) a function $u(z)$ which satisfies $u(0) = 0$ and $u'(0) = 0$. Hence, the manifold \mathcal{M} is also given by (5.2) with the function $\psi(z) = a + Qz + g'(a)^T u(z)$.



On the other hand, let \mathcal{M} be given by (5.2). Part (b) of the proof of Lemma 5.1 shows that $y = \psi(z)$ can be partitioned into $y_1 = \psi_1(z)$ and $y_2 = \psi_2(z)$, where ψ_1 is a local diffeomorphism. Consequently, \mathcal{M} is also given by (5.1) with $g(y) = y_2 - \psi_2(\psi_1^{-1}(y_1))$.

IV.5.2 Differential Equations on Manifolds

In Sect. IV.4 we introduced differential equations on a manifold as problems satisfying (4.2). With the help of Lemma 5.1 we are now in a position to characterize such problems without knowledge of the solutions.

Theorem 5.2. *Let \mathcal{M} be a submanifold of \mathbb{R}^n . The problem $\dot{y} = f(y)$ is a differential equation on the manifold \mathcal{M} (i.e., it satisfies (4.2)) if and only if*

$$f(y) \in T_y \mathcal{M} \quad \text{for all } y \in \mathcal{M}. \quad (5.6)$$

Proof. The necessity of (5.6) follows from the definition of $T_y \mathcal{M}$, because the exact solution of the differential equation lies in \mathcal{M} and has $f(y)$ as derivative.

To prove the sufficiency, we assume (5.6) and let \mathcal{M} be locally, near y_0 , be given by a parametrization $y = \psi(z)$ as in (5.2). We try to write the solution of $\dot{y} = f(y)$, $y(0) = y_0 = \psi(z_0)$ as $y(t) = \psi(z(t))$. If this is at all possible, then $z(t)$ must satisfy

$$\psi'(z)\dot{z} = f(\psi(z))$$

which, by assumption (5.6) and the second part of Lemma 5.1, is equivalent to

$$\dot{z} = \psi'(z)^+ f(\psi(z)), \quad (5.7)$$

where $A^+ = (A^T A)^{-1} A^T$ denotes the pseudo-inverse of a matrix with full column rank. Conversely, define $z(t)$ as the solution of (5.7) with $z(0) = z_0$, which is known to exist locally in t by the standard existence and uniqueness theory of ordinary differential equations on \mathbb{R}^m . Then $y(t) = \psi(z(t))$ is the solution of $\dot{y} = f(y)$ with $y(0) = y_0$. Hence, the solution $y(t)$ remains in \mathcal{M} . \square

We remark that the sufficiency proof of Theorem 5.2 only requires the function $f(y)$ to be defined on \mathcal{M} . Due to the equivalence of $\dot{y} = f(y)$ with (5.7) the problem is transported to the space of local coordinates. The standard local theory for ordinary differential equations on an Euclidean space (existence and uniqueness of solutions, . . .) can thus be extended in a straightforward way to differential equations on manifolds, i.e., $\dot{y} = f(y)$ with $f : \mathcal{M} \rightarrow \mathbb{R}^n$ satisfying (5.6).

IV.5.3 Numerical Integrators on Manifolds

Whereas the projection methods of Sect. IV.4 require the function $f(y)$ of the differential equation to be defined in a neighbourhood of \mathcal{M} (see Fig. 4.2), the numerical methods of this section evaluate $f(y)$ only on the manifold \mathcal{M} . The idea is to apply the numerical integrator in the parameter space rather than in the space where \mathcal{M} is embedded.

Algorithm 5.3 (Local Coordinates Approach). Assume that $y_n \in \mathcal{M}$ and that ψ is a local parametrization of \mathcal{M} satisfying $\psi(z_n) = y_n$. One step $y_n \mapsto y_{n+1}$ is defined as follows (see Fig. 5.1):

- Compute $\tilde{z}_{n+1} = \Phi_h(z_n)$, the result of the method Φ_h applied to (5.7);
- define the numerical solution by $y_{n+1} = \psi(\tilde{z}_{n+1})$.

It is important to remark that the parametrization $y = \psi(z)$ can be changed at every step.

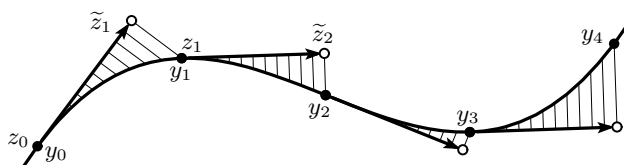


Fig. 5.1. The numerical solution of differential equations on manifolds via local coordinates

As indicated at the beginning of Sect. IV.5.1, there are many possible choices of local coordinates. Consider the pendulum equation of Example 4.1, where $\mathcal{M} = \{(q_1, q_2, p_1, p_2) \mid q_1^2 + q_2^2 = 1, q_1 p_1 + q_2 p_2 = 0\}$. A standard parametrization here is $q_1 = \sin \alpha$, $q_2 = -\cos \alpha$, $p_1 = \omega \cos \alpha$, and $p_2 = \omega \sin \alpha$. In the new coordinates (α, ω) the problem becomes simply $\dot{\alpha} = \omega$, $\dot{\omega} = -\sin \alpha$. Other typical choices are the exponential map $\psi(Z) = \exp(Z)$ for differential equations on Lie groups, and the Cayley transform $\psi(Z) = (I - Z)^{-1}(I + Z)$ for quadratic Lie groups. This will be studied in more detail in Sect. IV.8 below. Here we discuss two commonly used choices which do not use a special structure of the manifold.

Generalized Coordinate Partitioning. We assume that the manifold is given by (5.1). If $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ has a Jacobian with full rank m at $y = a$, we can find a partitioning $y = (y_1, y_2)$, such that $\partial g / \partial y_2(a)$ is invertible. In this case we can choose the components of y_1 as local coordinates. The function $y = \psi(z)$ is then given by $y_1 = z$ and $y_2 = \psi_2(z)$, where $\psi_2(z)$ is implicitly defined by $g(z, \psi_2(z)) = 0$. This approach has been promoted by Wehage & Haug (1982) in the context of constrained mechanical systems, and the partitioning is found by Gaussian elimination with full pivoting applied to the matrix $g'(a)$. Another way of finding the partitioning is by the use of the QR decomposition with column change.

Tangent Space Parametrization. Let the manifold \mathcal{M} be given by (5.1), and collect the vectors of an orthogonal basis of $T_a \mathcal{M}$ in the matrix Q . We then consider the parametrization

$$\psi_a(z) = a + Qz + g'(a)^T u(z), \quad (5.8)$$

where $u(z)$ is defined by $g(\psi_a(z)) = 0$, exactly as in the discussion after the proof of Lemma 5.1. Differentiating (5.8) yields

$$(Q + g'(a)^T u'(z))\dot{z} = \dot{y} = f(y) = f(\psi_a(z)).$$

Since $Q^T Q = I$ and $g'(a)Q = 0$, this relation is equivalent to the differential equation

$$\dot{z} = Q^T f(\psi_a(z)), \quad (5.9)$$

which corresponds to (5.7). If we apply a numerical method to (5.9), every function evaluation requires the projection of an element of the tangent space onto the manifold. This procedure is illustrated in Fig. 5.1, and was originally proposed by Potra & Rheinboldt (1991) for the solution of the Euler–Lagrange equations of constrained multibody systems (see also Hairer & Wanner (1996), p. 476).

IV.6 Differential Equations on Lie Groups

Theorem 1.6 and Lemma 3.1 are particular cases of a more general result which can be conveniently formulated with the concept of Lie groups and Lie algebras (see Olver (1986) and Varadarajan (1974) for an introduction to these subjects).

A *Lie group* is a group G which is a differentiable manifold, and for which the product is a differentiable mapping $G \times G \rightarrow G$. We restrict our considerations to *matrix Lie groups*, that is, Lie groups which are subgroups of $GL(n)$, the group of invertible $n \times n$ matrices with the usual matrix product as the group operation.

Example 6.1. An important example of a Lie group is the group

$$O(n) = \{Y \in GL(n) \mid Y^T Y = I\}$$

of all orthogonal matrices. It is the zero set of $g(Y) = Y^T Y - I$, where we consider g as a mapping from the set of all $n \times n$ matrices (i.e., $\mathbb{R}^{n \cdot n}$) to the set of all symmetric matrices (which can be identified with $\mathbb{R}^{n(n+1)/2}$). The derivative $g'(Y)$ is surjective for $Y \in O(n)$, because for any symmetric matrix K the choice $H = YK/2$ solves the equation $g'(Y)H = K$. Therefore, the matrix $g'(Y)$ has full rank (cf. (5.1)) so that $O(n)$ defines a differentiable manifold of dimension $n^2 - n(n+1)/2 = n(n-1)/2$. The set $O(n)$ is also a group with unit element I (the identity). Since the matrix multiplication is a differentiable mapping, $O(n)$ is a Lie group.

Table 6.1 lists further prominent examples. The matrix J appearing in the definition of the symplectic group is the matrix determining the symplectic structure on \mathbb{R}^n (see Sect. VI.2).

As the following lemma shows, the tangent space $\mathfrak{g} = T_I G$ at the identity I of a matrix Lie group G is closed under forming commutators of its elements. This makes \mathfrak{g} an algebra, the *Lie algebra* of the Lie group G .

Lemma 6.2 (Lie Bracket and Lie Algebra). *Let G be a matrix Lie group and let $\mathfrak{g} = T_I G$ be the tangent space at the identity. The Lie bracket (or commutator)*

$$[A, B] = AB - BA \quad (6.1)$$

defines an operation $\mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}$ which is bilinear, skew-symmetric ($[A, B] = -[B, A]$), and satisfies the Jacobi identity

$$[A, [B, C]] + [C, [A, B]] + [B, [C, A]] = 0. \quad (6.2)$$

³ Marius Sophus Lie, born: 17 December 1842 in Nordfjordeid (Norway), died: 18 February 1899.



Marius Sophus Lie³

Table 6.1. Some matrix Lie groups and their corresponding Lie algebras

Lie group	Lie algebra
$\mathrm{GL}(n) = \{Y \mid \det Y \neq 0\}$ general linear group	$\mathfrak{gl}(n) = \{A \mid \text{arbitrary matrix}\}$ Lie algebra of $n \times n$ matrices
$\mathrm{SL}(n) = \{Y \mid \det Y = 1\}$ special linear group	$\mathfrak{sl}(n) = \{A \mid \mathrm{trace}(A) = 0\}$ special linear Lie algebra
$\mathrm{O}(n) = \{Y \mid Y^T Y = I\}$ orthogonal group	$\mathfrak{so}(n) = \{A \mid A^T + A = 0\}$ skew-symmetric matrices
$\mathrm{SO}(n) = \{Y \in \mathrm{O}(n) \mid \det Y = 1\}$ special orthogonal group	$\mathfrak{so}(n) = \{A \mid A^T + A = 0\}$ skew-symmetric matrices
$\mathrm{Sp}(n) = \{Y \mid Y^T J Y = J\}$ symplectic group	$\mathfrak{sp}(n) = \{A \mid JA + A^T J = 0\}$

Proof. By definition of the tangent space, for $A, B \in \mathfrak{g}$, there exist differentiable paths $\alpha(t), \beta(t)$ ($|t| < \varepsilon$) in G such that $\alpha(t) = I + tA(t)$ with a continuous function $A(t)$ with $A(0) = A$, and similarly $\beta(t) = I + tB(t)$ with $B(0) = B$. Now consider the path $\gamma(t)$ in G defined by

$$\gamma(t) = \alpha(\sqrt{t})\beta(\sqrt{t})\alpha(\sqrt{t})^{-1}\beta(\sqrt{t})^{-1}, \quad t \geq 0.$$

An elementary computation then yields

$$\gamma(t) = I + t[A, B] + o(t).$$

With the extension $\gamma(t) = \gamma(-t)^{-1}$ for negative t , this is a differentiable path in G satisfying $\gamma(0) = I$ and $\dot{\gamma}(0) = [A, B]$. Hence $[A, B] \in \mathfrak{g}$ by definition of the tangent space. The properties of the Lie bracket can be verified in a straightforward way. \square

Example 6.3. Consider again the orthogonal group $\mathrm{O}(n)$. Since the derivative of $g(Y) = Y^T Y - I$ at the identity is $g'(I)H = I^T H + H^T I = H + H^T$, it follows from the first part of Lemma 5.1 that the Lie algebra corresponding to $\mathrm{O}(n)$ consists of all skew-symmetric matrices. The right column of Table 6.1 gives the Lie algebras of the other Lie groups listed there.

The following basic lemma shows that the exponential map yields a local parametrization of the Lie group near the identity, with the Lie algebra (a linear space) as the parameter space.

Lemma 6.4 (Exponential Map). *Consider a matrix Lie group G and its Lie algebra \mathfrak{g} . The matrix exponential is a map*

$$\exp : \mathfrak{g} \rightarrow G,$$

i.e., for $A \in \mathfrak{g}$ we have $\exp(A) \in G$. Moreover, \exp is a local diffeomorphism in a neighbourhood of $A = 0$.

Proof. For $A \in \mathfrak{g}$, it follows from the definition of the tangent space $\mathfrak{g} = T_I G$ that there exists a differentiable path $\alpha(t)$ in G satisfying $\alpha(0) = I$ and $\dot{\alpha}(0) = A$. For a fixed $Y \in G$, the path $\gamma(t) := \alpha(t)Y$ is in G and satisfies $\gamma(0) = Y$ and $\dot{\gamma}(0) = AY$. Consequently, $AY \in T_Y G$ and $\dot{Y} = AY$ defines a differential equation on the manifold G . The solution $Y(t) = \exp(tA)$ is therefore in G for all t .

Since $\exp(H) - \exp(0) = H + \mathcal{O}(H^2)$, the derivative of the exponential map at $A = 0$ is the identity, and it follows from the inverse function theorem that \exp is a local diffeomorphism close to $A = 0$. \square

The proof of Lemma 6.4 shows that for a matrix Lie group G the tangent space at $Y \in G$ has the form

$$T_Y G = \{AY \mid A \in \mathfrak{g}\}. \quad (6.3)$$

By Theorem 5.2, differential equations on a matrix Lie group (considered as a manifold) can therefore be written as

$$\dot{Y} = A(Y)Y \quad (6.4)$$

where $A(Y) \in \mathfrak{g}$ for all $Y \in G$. The following theorem summarizes this discussion, and extends the statements of Theorem 1.6 and Lemma 3.1 to more general matrix Lie groups.

Theorem 6.5. *Let G be a matrix Lie group and \mathfrak{g} its Lie algebra. If $A(Y) \in \mathfrak{g}$ for all $Y \in G$ and if $Y_0 \in G$, then the solution of (6.4) satisfies $Y(t) \in G$ for all t . \square*

If in addition $A(Y) \in \mathfrak{g}$ for all matrices Y , and if

$$G = \{Y \mid g(Y) = \text{Const}\}$$

is one of the Lie groups of Table 6.1, then $g(Y)$ is an invariant of the differential equation (6.4) in the sense of Definition 1.1.

IV.7 Methods Based on the Magnus Series Expansion



Wilhelm Magnus⁴

Before we discuss the numerical solution of differential equations (6.4) on Lie groups, let us give an explicit formula for the solution of linear matrix differential equations

$$\dot{Y} = A(t)Y. \quad (7.1)$$

No assumption on the matrix $A(t)$ is made for the moment (apart from continuous dependence on t). For the scalar case, the solution of (7.1) with $Y(0) = Y_0$ is given by

$$Y(t) = \exp\left(\int_0^t A(\tau) d\tau\right) Y_0. \quad (7.2)$$

Also in the case where the matrices $A(t)$ and $\int_0^t A(\tau) d\tau$ commute, (7.2) is the solution of (7.1). In the general non-commutative case

we follow the approach of Magnus (1954) and we search for a matrix function $\Omega(t)$ such that

$$Y(t) = \exp(\Omega(t)) Y_0$$

solves (7.1). The main ingredient for the solution will be the inverse of the derivative of the matrix exponential. It has been studied in Sect. III.4, Lemma III.4.2, and is given by

$$d \exp_{\Omega}^{-1}(H) = \sum_{k \geq 0} \frac{B_k}{k!} \operatorname{ad}_{\Omega}^k(H), \quad (7.3)$$

where B_k are the Bernoulli numbers, and $\operatorname{ad}_{\Omega}(A) = [\Omega, A] = \Omega A - A \Omega$ is the adjoint operator introduced in (III.4.1).

Theorem 7.1 (Magnus 1954). *The solution of the differential equation (7.1) can be written as $Y(t) = \exp(\Omega(t)) Y_0$ with $\Omega(t)$ defined by*

$$\dot{\Omega} = d \exp_{\Omega}^{-1}(A(t)), \quad \Omega(0) = 0. \quad (7.4)$$

As long as $\|\Omega(t)\| < \pi$, the convergence of the $d \exp_{\Omega}^{-1}$ expansion (7.3) is assured.

Proof. Comparing the derivative of $Y(t) = \exp(\Omega(t)) Y_0$,

$$\dot{Y}(t) = \left(\frac{d}{d\Omega} \exp \Omega(t) \right) \dot{\Omega}(t) Y_0 = \left(d \exp_{\Omega(t)}(\dot{\Omega}(t)) \right) \exp(\Omega(t)) Y_0,$$

with (7.1) we obtain $A(t) = d \exp_{\Omega(t)}(\dot{\Omega}(t))$. Applying the inverse operator $d \exp_{\Omega}^{-1}$ to this relation yields the differential equation (7.4) for $\Omega(t)$. The statement on the convergence is a consequence of Lemma III.4.2. \square

⁴ Wilhelm Magnus, born: 5 February 1907 in Berlin (Germany), died: 15 October 1990.

The first few Bernoulli numbers are $B_0 = 1$, $B_1 = -1/2$, $B_2 = 1/6$, $B_3 = 0$. The differential equation (7.4) therefore becomes

$$\dot{\Omega} = A(t) - \frac{1}{2} [\Omega, A(t)] + \frac{1}{12} [\Omega, [\Omega, A(t)]] + \dots,$$

which is nonlinear in Ω . Applying Picard fixed point iteration after integration yields

$$\begin{aligned} \Omega(t) = & \int_0^t A(\tau) d\tau - \frac{1}{2} \int_0^t \left[\int_0^\tau A(\sigma) d\sigma, A(\tau) \right] d\tau \\ & + \frac{1}{4} \int_0^t \left[\int_0^\tau \left[\int_0^\sigma A(\mu) d\mu, A(\sigma) \right] d\sigma, A(\tau) \right] d\tau \quad (7.5) \\ & + \frac{1}{12} \int_0^t \left[\int_0^\tau A(\sigma) d\sigma, \left[\int_0^\tau A(\mu) d\mu, A(\tau) \right] \right] d\tau + \dots, \end{aligned}$$

which is the so-called *Magnus expansion*. For smooth matrices $A(t)$ the remainder in (7.5) is of size $\mathcal{O}(t^5)$ so that the truncated series inserted into $Y(t) = \exp(\Omega(t))Y_0$ gives an excellent approximation to the solution of (7.1) for small t .

Numerical Methods Based on the Magnus Expansion. Iserles & Nørsett (1999) study the general form of the Magnus expansion (7.5), and they relate the iterated integrals and the rational coefficients in (7.5) to binary trees. For a numerical integration of

$$\dot{Y} = A(t)Y, \quad Y(t_0) = Y_0 \quad (7.6)$$

(where Y is a matrix or a vector) they propose using $Y_{n+1} = \exp(h\Omega_n)Y_n$, where $h\Omega_n$ is a suitable approximation of $\Omega(h)$ given by (7.5) with $A(t_n + \tau)$ instead of $A(\tau)$. Of course, the Magnus expansion has to be truncated and the integrals have to be approximated by numerical quadrature.

We follow here the collocation approach suggested by Zanna (1999). The idea is to replace $A(t)$ locally by an interpolation polynomial

$$\hat{A}(t) = \sum_{i=1}^s \ell_i(t) A(t_n + c_i h),$$

and to solve $\dot{Y} = \hat{A}(t)Y$ on $[t_n, t_n + h]$ by the use of the truncated series (7.5).

Theorem 7.2. Consider a quadrature formula $(b_i, c_i)_{i=1}^s$ of order $p \geq s$, and let $Y(t)$ and $Z(t)$ be solutions of $\dot{Y} = A(t)Y$ and $\dot{Z} = \hat{A}(t)Z$, respectively, satisfying $Y(t_n) = Z(t_n)$. Then, $Z(t_n + h) - Y(t_n + h) = \mathcal{O}(h^{p+1})$.

Proof. We write the differential equation for Z as $\dot{Z} = A(t)Z + (\hat{A}(t) - A(t))Z$ and use the variation of constants formula to get

$$Z(t_n + h) - Y(t_n + h) = \int_{t_n}^{t_n+h} R(t_n + h, \tau) (\hat{A}(\tau) - A(\tau)) Z(\tau) d\tau.$$

Applying our quadrature formula to this integral gives zero as result, and the remainder is of size $\mathcal{O}(h^{p+1})$. Details of the proof are as for Theorem II.1.5. \square

Example 7.3. As a first example, we use the midpoint rule ($c_1 = 1/2$, $b_1 = 1$). In this case the interpolation polynomial is constant, and the method becomes

$$Y_{n+1} = \exp\left(hA(t_n + h/2)\right) Y_n, \quad (7.7)$$

which is of order 2.

Example 7.4. The two-stage Gauss quadrature is given by $c_{1,2} = 1/2 \pm \sqrt{3}/6$, $b_{1,2} = 1/2$. The interpolation polynomial is of degree one and we have to apply (7.5) in order to get an approximation Y_{n+1} . Since we are interested in a fourth order approximation, we can neglect the remainder term (indicated by \dots in (7.5)). Computing analytically the iterated integrals over products of $\ell_i(t)$ we obtain

$$Y_{n+1} = \exp\left(\frac{h}{2}(A_1 + A_2) + \frac{\sqrt{3}h^2}{12}[A_2, A_1]\right) Y_n, \quad (7.8)$$

where $A_1 = A(t_n + c_1 h)$ and $A_2 = A(t_n + c_2 h)$. This is a method of order four. The terms of (7.5) with triple integrals give $\mathcal{O}(h^4)$ expressions, whose leading term vanishes by the symmetry of the method (Exercise V.7). Therefore, they need not be considered.

Theorem 7.2 allows us to obtain methods of arbitrarily high order. A straightforward use of the expansion (7.5) yields an expression with a large number of commutators. Munthe-Kaas & Owren (1999) and Blanes, Casas & Ros (2000a) construct higher order methods with a reduced number of commutators. For example, for order 6 the required number of commutators is reduced from 7 to 4.

Let us remark that all numerical methods of this section are of the form $Y_{n+1} = \exp(h\Omega_n)Y_n$, where Ω_n is a linear combination of $A(t_n + c_i h)$ and of their commutators. If $A(t) \in \mathfrak{g}$ for all t , then also $h\Omega_n$ lies in the Lie algebra \mathfrak{g} , so that the numerical solution stays in the Lie group G if $Y_0 \in G$ (this is a consequence of Lemma 6.4).

IV.8 Lie Group Methods

Consider a differential equation

$$\dot{Y} = A(Y)Y, \quad Y(0) = Y_0 \quad (8.1)$$

on a matrix Lie group G . This means that $Y_0 \in G$ and that $A(Y) \in \mathfrak{g}$ for all $Y \in G$. Since this is a special case of differential equations on a manifold, projection methods (Sect. IV.4) as well as methods based on local coordinates (Sect. IV.5) are well suited for their numerical treatment. Here we present further approaches which also yield approximations that lie on the manifold.

All numerical methods of this section can be extended in a straightforward way to non-autonomous problems $\dot{Y} = A(t, Y)Y$ with $A(t, Y) \in \mathfrak{g}$ for all t and all $Y \in G$. Just to simplify the notation we restrict ourselves to the formulation (8.1).

IV.8.1 Crouch-Grossman Methods

The discipline of Lie-group methods owes a great deal to the pioneering work of Peter Crouch and his co-workers . . .

(A. Iserles, H.Z. Munthe-Kaas, S.P. Nørsett & A. Zanna 2000)

The numerical approximation of explicit Runge–Kutta methods is obtained by a composition of the following two basic operations: (i) an evaluation of the vector field $f(Y) = A(Y)Y$ and (ii) a computation of an update of the form $Y + hf(Z)$. For example, the left method of (II.1.3) consists of the following steps: evaluate $K_1 = f(Y_0)$; compute $\tilde{Y}_1 = Y_0 + hK_1$; evaluate $K_2 = f(\tilde{Y}_1)$; compute $Y_{1/2} = Y_0 + \frac{h}{2}K_1$; compute $Y_1 = Y_{1/2} + \frac{h}{2}K_2$.

In the context of differential equations on Lie groups, these methods have the disadvantage that, even when $Y \in G$ and $Z \in G$, the update $Y + hA(Z)Z$ is in general not in the Lie group. The idea of Crouch & Grossman (1993) is to replace the “update” operation with $\exp(hA(Z))Y$.

Definition 8.1. Let b_i, a_{ij} ($i, j = 1, \dots, s$) be real numbers. An explicit s -stage Crouch-Grossman method is given by

$$\begin{aligned} Y^{(i)} &= \exp(ha_{i,i-1}K_{i-1}) \cdots \exp(ha_{i1}K_1)Y_n, & K_i &= A(Y^{(i)}), \\ Y_{n+1} &= \exp(hb_sK_s) \cdots \exp(hb_1K_1)Y_n. \end{aligned}$$

For example, the method of Runge described above ($s = 2$, $a_{21} = 1$, $b_1 = b_2 = 1/2$) leads to

$$Y_{n+1} = \exp\left(\frac{h}{2}K_2\right) \exp\left(\frac{h}{2}K_1\right)Y_n, \quad (8.2)$$

where $K_1 = A(Y_n)$ and $K_2 = A(\exp(hK_1)Y_n)$.

By construction, the methods of Crouch-Grossman give rise to approximations Y_n which lie exactly on the manifold defined by the Lie group. But what can be said about their order of accuracy?

Theorem 8.2. Let $c_i = \sum_j a_{ij}$. A Crouch-Grossman method has order p ($p \leq 3$) if the following order conditions are satisfied:

$$\text{order } 1: \quad \sum_i b_i = 1 \quad (8.3)$$

$$\text{order } 2: \quad \sum_i b_i c_i = 1/2 \quad (8.4)$$

$$\text{order } 3: \quad \sum_i b_i c_i^2 = 1/3 \quad (8.5)$$

$$\sum_{ij} b_i a_{ij} c_j = 1/6 \quad (8.6)$$

$$\sum_i b_i^2 c_i + 2 \sum_{i < j} b_i c_i b_j = 1/3. \quad (8.7)$$

Proof. As in the case of Runge–Kutta methods, the order conditions can be found by comparing the Taylor series expansions of the exact and the numerical solution. In addition to the conditions stated in the theorem, this leads to relations such as

$$\sum_i b_i^2 c_i + 2 \sum_{i < j} b_i b_j c_j = \frac{2}{3}. \quad (8.8)$$

Adding this equation to (8.7) we find $2 \sum_{i,j} b_i c_i b_j = 1$, which is satisfied by (8.3) and (8.4). Hence, the relation (8.8) is already a consequence of the conditions stated in the theorem. \square

Table 8.1. Crouch-Grossman methods of order 3

0				0			
-1/24	-1/24			3/4	3/4		
17/24	161/24	-6		17/24	119/216	17/108	
	1	-2/3	2/3		13/51	-2/3	24/17

Crouch & Grossman (1993) present several solutions of the system (8.3)–(8.7), one of which is given in the left array of Table 8.1. The construction of higher order Crouch-Grossman methods is very complicated (“... any attempt to analyze algorithms of order greater than three will be very complex, ...”, Crouch & Grossman, 1993).

The theory of order conditions for Runge–Kutta methods (Sect. III.1) has been extended to Crouch-Grossman methods by Owren & Marthinsen (1999). It turns out that the order conditions for classical Runge–Kutta methods form a subset of those for Crouch-Grossman methods. The first new condition is (8.7). For a method of order 4, thirteen conditions (including those of Theorem 8.2) have to be satisfied. Solving these equations, Owren & Marthinsen (1999) construct a 4th order method with $s = 5$ stages.

IV.8.2 Munthe-Kaas Methods

These methods were developed in a series of papers by Munthe-Kaas (1995, 1998, 1999). The main motivation behind the first of these papers was to develop a theory of Runge–Kutta methods in a coordinate-free framework. After attempts that led to new order conditions (as for the Crouch-Grossman methods), Munthe-Kaas (1999) had the idea to write the solution as $Y(t) = \exp(\Omega(t))Y_0$ and to solve numerically the differential equation for $\Omega(t)$. It sounds awkward to replace the differential equation (8.1) by a more complicated one. However, the nonlinear invariants $g(Y) = 0$ of (8.1) defining the Lie group are replaced with linear invariants $g'(I)(\Omega) = 0$ defining the Lie algebra, and we know from Sect. IV.1 that essentially all numerical methods automatically conserve linear invariants.

It follows from the proof of Theorem 7.1 that the solution of (8.1) can be written as $Y(t) = \exp(\Omega(t))Y_0$, where $\Omega(t)$ is the solution of $\dot{\Omega} = d\exp_{\Omega}^{-1}(A(Y(t)))$, $\Omega(0) = 0$. Since it is not practical to work with the operator $d\exp_{\Omega}^{-1}$, we truncate the series (7.3) suitably and consider the differential equation

$$\dot{\Omega} = A(\exp(\Omega)Y_0) + \sum_{k=1}^q \frac{B_k}{k!} \operatorname{ad}_{\Omega}^k \left(A(\exp(\Omega)Y_0) \right), \quad \Omega(0) = 0. \quad (8.9)$$

This leads to the following method.

Algorithm 8.3 (Munthe-Kaas 1999). *Consider the problem (8.1) with $A(Y) \in \mathfrak{g}$ for $Y \in G$. Assume that Y_n lies in the Lie group G . Then, the step $Y_n \mapsto Y_{n+1}$ is defined as follows:*

- *consider the differential equation (8.9) with Y_n instead of Y_0 , and apply a Runge–Kutta method (explicit or implicit) to get an approximation $\Omega_1 \approx \Omega(h)$,*
- *then define the numerical solution by $Y_{n+1} = \exp(\Omega_1)Y_n$.*

Before analyzing this algorithm, we emphasize its close relationship with Algorithm 5.3. In fact, if we identify the Lie algebra \mathfrak{g} with \mathbb{R}^k (where k is the dimension of the vector space \mathfrak{g}), the mapping $\psi(\Omega) = \exp(\Omega)Y_n$ is a local parametrization of the Lie group G (see Lemma 6.4). Apart from the truncation of the series in (8.9), Algorithm 8.3 is a special case of Algorithm 5.3.

Important properties of the Munthe-Kaas methods are given in the next two theorems.

Theorem 8.4. *Let G be a matrix Lie group and \mathfrak{g} its Lie algebra. If $A(Y) \in \mathfrak{g}$ for $Y \in G$ and if $Y_0 \in G$, then the numerical solution of the Lie group method of Algorithm 8.3 lies in G , i.e., $Y_n \in G$ for all $n = 0, 1, 2, \dots$.*

Proof. It is sufficient to prove that for $Y_0 \in G$ the numerical solution Ω_1 of the Runge–Kutta method applied to (8.9) lies in \mathfrak{g} . Since the Lie bracket $[\Omega, A]$ is an operation $\mathfrak{g} \times \mathfrak{g} \rightarrow \mathfrak{g}$, and since $\exp(\Omega)Y_0 \in G$ for $\Omega \in \mathfrak{g}$, the right-hand expression of (8.9) is in \mathfrak{g} for $\Omega \in \mathfrak{g}$. Hence, (8.9) is a differential equation on the vector space \mathfrak{g} with solution $\Omega(t) \in \mathfrak{g}$. All operations in a Runge–Kutta method give results in \mathfrak{g} , so that the numerical approximation Ω_1 also lies in \mathfrak{g} . \square

Theorem 8.5. *If the Runge–Kutta method is of (classical) order p and if the truncation index in (8.9) satisfies $q \geq p - 2$, then the method of Algorithm 8.3 is of order p .*

Proof. For sufficiently smooth $A(Y)$ we have $\Omega(t) = tA(Y_0) + \mathcal{O}(t^2)$, $Y(t) = Y_0 + \mathcal{O}(t)$ and $[\Omega(t), A(Y(t))] = \mathcal{O}(t^2)$. This implies that $\operatorname{ad}_{\Omega(t)}^k(A(Y(t))) = \mathcal{O}(t^{k+1})$, so that the truncation of the series in (8.9) induces an error of size $\mathcal{O}(h^{q+2})$ for $|t| \leq h$. Hence, for $q + 2 \geq p$, this truncation does not affect the order of convergence. \square

The most simple Lie group method is obtained if we take the explicit Euler method as basic discretization and $q = 0$ in (8.9). This leads to the so-called *Lie–Euler method*

$$Y_{n+1} = \exp(hA(Y_n))Y_n. \quad (8.10)$$

This is also a special case of the Crouch-Grossman methods of Definition 8.1.

Taking the implicit midpoint rule as the basic discretization and again $q = 0$ in (8.9), we obtain the *Lie midpoint rule*

$$Y_{n+1} = \exp(\Omega)Y_n, \quad \Omega = hA(\exp(\Omega/2)Y_n). \quad (8.11)$$

This is an implicit equation in Ω and has to be solved by fixed point iteration or by Newton-type methods.

Example 8.6. We take the coefficients of the right array of Table 8.1. They give rise to 3rd order Munthe-Kaas and 3rd order Crouch-Grossman methods. We apply both methods with the large step size $h = 0.35$ to the system (1.5) which is already of the form (8.1). Observe that Y_0 is a vector in \mathbb{R}^3 and not a matrix, but all results of this section remain valid for this case. For the computation of the matrix exponential we use the Rodrigues formula (Exercise 17). The numerical results (first 1000 steps) are shown in Fig. 8.1. We see that the numerical solution stays on the manifold (sphere), but on the sphere the qualitative behaviour is not correct. A similar behaviour could be observed for projection methods (the orthogonal projection consists simply in dividing the approximation \tilde{Y}_{n+1} by its norm) and by the methods based on local coordinates.

Crouch-Grossman methods and Munthe-Kaas methods are very similar. If they are based on the same set of Runge-Kutta coefficients, both methods use s evaluations of the matrix $A(Y)$. The Crouch-Grossman methods require in general the computation of $s(s+1)/2$ matrix exponentials, whereas the Munthe-Kaas methods require only s of them. On the other hand, Munthe-Kaas methods need also the computations of a certain number of commutators which increases with q in (8.9). In such a comparison one has to take into account that every classical Runge-Kutta method defines a Munthe-Kaas method of the same order, but Crouch-Grossman methods of high order are very difficult to obtain, and need more stages for the same order (if $p \geq 4$).

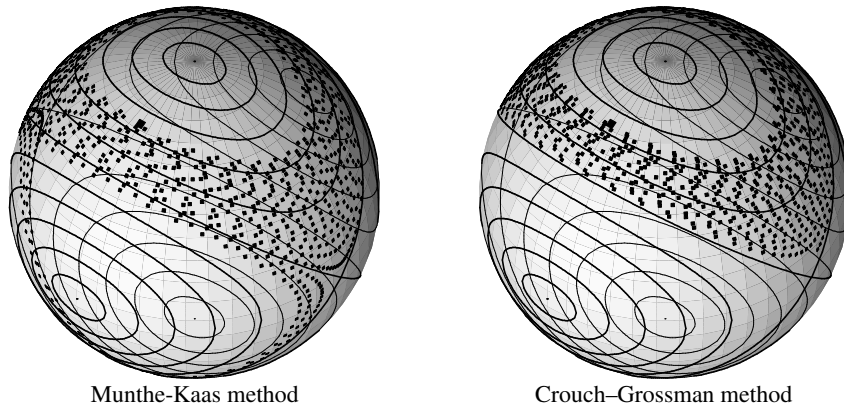


Fig. 8.1. Solutions of the Euler equations (1.4) for the rigid body

IV.8.3 Further Coordinate Mappings

The methods of Algorithm 8.3 are based on the local parametrization $\psi(\Omega) = \exp(\Omega)Y_n$. For all Lie groups, this is a diffeomorphism between the Lie group and the corresponding Lie algebra. Are there other, computationally more efficient parametrizations that can be used in special situations?

The Cayley Transform. Lie groups of the form

$$G = \{Y \mid Y^T P Y = P\}, \quad (8.12)$$

where P is a given constant matrix, are called *quadratic Lie groups*. The corresponding Lie algebra is given by $\mathfrak{g} = \{\Omega \mid P\Omega + \Omega^T P = 0\}$. The orthogonal group $O(n)$ and the symplectic group $Sp(n)$ are prominent special cases (see Table 6.1). For such groups we have the following analogue of Lemma 6.4.

Lemma 8.7. *For a quadratic Lie group G , the Cayley transform*

$$\text{cay } \Omega = (I - \Omega)^{-1}(I + \Omega)$$

maps elements of \mathfrak{g} into G . Moreover, it is a local diffeomorphism near $\Omega = 0$.

Proof. For $\Omega \in \mathfrak{g}$ (i.e., $P\Omega + \Omega^T P = 0$) we have $P(I + \Omega) = (I - \Omega)^T P$ and also $P(I - \Omega)^{-1} = (I + \Omega)^{-T} P$. For $Y = (I - \Omega)^{-1}(I + \Omega)$ this immediately implies $Y^T P Y = P$. \square

The use of the Cayley transform for the numerical integration of differential equations on Lie groups has been proposed by Lewis & Simo (1994) and Diele, Lopez & Peluso (1998) for the orthogonal group, and by Lopez & Politi (2001) for general quadratic groups. It is based on the following result, which is an adaptation of Lemma III.4.1 and Lemma III.4.2 to the Cayley transform.

Lemma 8.8. *The derivative of $\text{cay } \Omega$ is given by*

$$\left(\frac{d}{d\Omega} \text{cay } \Omega\right)H = \left(d \text{cay } \Omega(H)\right) \text{cay } \Omega,$$

where

$$d \text{cay } \Omega(H) = 2(I - \Omega)^{-1}H(I + \Omega)^{-1}. \quad (8.13)$$

For the inverse of $d \text{cay } \Omega$ we have

$$d \text{cay } \Omega^{-1}(H) = \frac{1}{2}(I - \Omega)H(I + \Omega). \quad (8.14)$$

Proof. By the usual rules of calculus we obtain

$$\left(\frac{d}{d\Omega} \text{cay } \Omega\right)H = (I - \Omega)^{-1}H(I - \Omega)^{-1}(I + \Omega) + (I - \Omega)^{-1}H,$$

and a simple algebraic manipulation proves the statements. \square

The numerical approach for solving (8.1) in the case of quadratic Lie groups is an adaptation of the Algorithm 8.3. We consider the local parametrization $Y = \psi(\Omega) = \text{cay}(\Omega)Y_n$, and we apply one step of a numerical method to the differential equation $\dot{\Omega} = d \text{cay}_{\Omega}^{-1} A(\text{cay}(\Omega)Y_n)$ which, by (8.14), is equivalent to

$$\dot{\Omega} = \frac{1}{2}(I - \Omega)A(\text{cay}(\Omega)Y_n)(I + \Omega).$$

This equation replaces (8.9) in the Algorithm 8.3. Since no truncation of an infinite series is necessary here, this approach is a special case of Algorithm 5.3.

Canonical Coordinates of the Second Kind. For a basis $\{C_1, C_2, \dots, C_d\}$ of the Lie algebra \mathfrak{g} the coordinates z_1, \dots, z_d of the local parametrization $\psi(z) = \exp(\sum_{i=1}^d z_i C_i)$ of the Lie group G are called *canonical coordinates of the first kind*. Here we are interested in the parametrization

$$\psi(z) = \exp(z_1 C_1) \exp(z_2 C_2) \cdots \exp(z_d C_d), \quad (8.15)$$

and we call $z = (z_1, \dots, z_d)^T$ *canonical coordinates of the second kind* (Varadarajan 1974). The use of these coordinates in connection with the numerical solution of differential equations on Lie groups has been promoted by Celledoni & Iserles (2001) and Owren & Marthinsen (2001). The idea behind this choice is that, due to a sparse structure of the C_i , the computation of $\exp(z_1 C_1), \dots, \exp(z_d C_d)$ may be much cheaper than the computation of $\exp(\sum_i z_i C_i)$.

With the change of coordinates $y = \psi(z)$, the differential equation (8.1) becomes $\psi'(z)\dot{z} = A(\psi(z))\psi(z)$, which is equivalent to

$$\begin{aligned} A(\psi(z)) &= \sum_{i=1}^d \dot{z}_i \exp(z_1 C_1) \cdots \exp(z_{i-1} C_{i-1}) \\ &\quad \cdot C_i \cdot \exp(-z_{i-1} C_{i-1}) \cdots \exp(-z_1 C_1) \\ &= \sum_{i=1}^d \dot{z}_i (F_1 \circ \cdots \circ F_{i-1}) C_i, \end{aligned} \quad (8.16)$$

where we use the notation $F_j C = \exp(z_j C_j) C \exp(-z_j C_j)$ for the linear operator $F_j : \mathfrak{g} \rightarrow \mathfrak{g}$; see Exercise 12. We need to compute $\dot{z}_1, \dots, \dot{z}_d$ from (8.16), and this will usually be a computationally expensive task. However, for several Lie algebras and for well chosen bases this can be done very efficiently. The crucial idea is the following: we let \hat{F}_j be defined by

$$\hat{F}_j C_i = \begin{cases} F_j C_i & \text{if } i > j \\ C_i & \text{if } i \leq j, \end{cases} \quad (8.17)$$

and we assume that

$$(F_1 \circ \cdots \circ F_{i-1}) C_i = (\hat{F}_1 \circ \cdots \circ \hat{F}_{i-1}) C_i, \quad i = 2, \dots, d. \quad (8.18)$$

Under this assumption, we have $(F_1 \circ \dots \circ F_{i-1})C_i = (\hat{F}_1 \circ \dots \circ \hat{F}_{i-1})C_i = (\hat{F}_1 \circ \dots \circ \hat{F}_{d-1})C_i$, and the relation (8.16) becomes

$$(\hat{F}_1 \circ \dots \circ \hat{F}_{d-1}) \left(\sum_{i=1}^d \dot{z}_i C_i \right) = A(\psi(z)). \quad (8.19)$$

In the situations which we have in mind, the operators \hat{F}_j can be efficiently inverted, and Algorithm 5.3 can be applied to the solution of (8.1).

The main difficulty of using this coordinate transform is to find a suitable ordering of a basis such that condition (8.18) is satisfied. The following lemma simplifies this task. We use the notation $\alpha_k(C)$ for the coefficient in the representation $C = \sum_{k=1}^d \alpha_k(C) C_k$.

Lemma 8.9. *Let $\{C_1, \dots, C_d\}$ be a basis of the Lie algebra \mathfrak{g} . If for every pair $j < i$ and for $k < j$ we have*

$$\alpha_k(F_j C_i) \neq 0 \quad \implies \quad F_\ell C_k = C_k \quad \text{for } \ell \text{ satisfying } k \leq \ell < j, \quad (8.20)$$

then the relation (8.18) holds for all $i = 2, \dots, d$.

Proof. We write $\hat{F}_{i-1} C_i = F_{i-1} C_i = \sum_k \alpha_k(F_{i-1} C_i) C_k$. It follows from the definition of \hat{F}_j and from (8.20) that $(\hat{F}_{i-2} \circ \hat{F}_{i-1}) C_i = (F_{i-2} \circ F_{i-1}) C_i$. A repeated application of this argument proves the statement. \square

Owren & Marthinsen (2001) have studied Lie algebras that admit a basis satisfying (8.18) for all z . We present here one of their examples.

Example 8.10 (Special Linear Group). Consider the differential equation (8.1) on the Lie group $SL(n) = \{Y \mid \det Y = 1\}$, i.e., the matrix $A(Y)$ lies in $\mathfrak{sl}(n) = \{A \mid \text{trace } A = 0\}$. As a basis of the Lie algebra $\mathfrak{sl}(n)$ we choose $E_{ij} = e_i e_j^T$ for $i \neq j$, and $D_i = e_i e_i^T - e_{i+1} e_{i+1}^T$ for $1 \leq i < n$ (here, $e_i = (0, \dots, 1, \dots, 0)^T$ denotes the vector whose only non-zero element is in the i th position). Following Owren & Marthinsen (2001) we order the elements of this basis as

$$\begin{aligned} E_{12} &< \dots < E_{1n} < E_{23} < \dots < E_{2n} < \dots < E_{n-1,n} \\ &< E_{21} < \dots < E_{n1} < E_{32} < \dots < E_{n2} < \dots < E_{n,n-1} \\ &< D_1 < \dots < D_{n-1}. \end{aligned}$$

With the use of Lemma 8.9 one can check in a straightforward way that the relation (8.18) is satisfied. In nearly all situations $\alpha_k(F_j C_i) = 0$ for $k < j < i$, so that (8.18) represents an empty condition. Consequently, the \dot{z}_i can be computed from (8.19). Due to the sparsity of the matrices E_{ij} and D_i , the computation of \hat{F}_i^{-1} can be done very efficiently.

IV.9 Geometric Numerical Integration Meets Geometric Numerical Linear Algebra

The persistent use of orthogonal transformations is a hallmark of numerical linear algebra. Correspondingly, manifolds incorporating orthogonality constraints play an important role all over this field; see Edelman, Arias & Smith (1998) on the geometry of algorithms with orthogonality constraints. In addition to the orthogonal group $O(n)$, the manifolds of primary interest are:

- $\mathcal{V}_{n,k}$, the Stiefel manifold of $n \times k$ matrices with k orthonormal columns,
- $\mathcal{G}_{n,k}$, the Grassmann manifold of orthogonal projections of \mathbb{R}^n onto k -dimensional subspaces, and
- $\mathcal{M}_k^{m \times n}$, the manifold of $m \times n$ matrices of rank k , which is related to orthogonal transformations via the singular value decomposition and a related decomposition discussed below.

IV.9.1 Numerical Integration on the Stiefel Manifold

The original motivation for Stiefel Manifolds (in Stiefel 1935) was the topological problem, whether a manifold \mathcal{M} can possess k everywhere linearly independent continuous vector fields. The problem, which had been solved for the case $k = 1$, was much harder for $k > 1$. In order to attack this question, Stiefel introduced ‘his’ manifold

$$\mathcal{V}_{n,k} = \{Y \in \mathbb{R}^{n \times k} \mid Y^T Y = I\}, \quad (9.1)$$

as an auxiliary tool for the definition of what later became known as the Stiefel-Whitney classes⁶.

Here, we are interested in computations on these manifolds for their own, with many applications, as for example the computation of Lyapunov exponents of differential equations; see Exercise 22 as well as Bridges & Reich (2001) and Dieci, Russell & van Vleck (1997). There are also many cases where orthogonality constraints concern only some of the variables in a differential equation. In molecular dynamics, for example, such orthogonality constraints arise in the Car-Parrinello approach to *ab initio* molecular dynamics (Car & Parrinello 1985) and in the multiconfiguration time-dependent Hartree method of quantum molecular dynamics (Beck, Jäckle, Worth & Meyer 2000).



Eduard Stiefel⁵

⁵ Eduard L. Stiefel, born: 21 April 1909 in Zürich, died: 25 November 1979; photo: Bildarchiv ETH-Bibliothek, Zürich.

⁶ We are grateful to our colleague A. Haeffliger for this indication.

Tangent and Normal Space. We choose a fixed matrix Y in the Stiefel manifold $\mathcal{V} = \mathcal{V}_{n,k}$. Then the tangent space (5.4) at $Y \in \mathcal{V}$ consists of the matrices Z such that $(Y + \varepsilon Z)^T(Y + \varepsilon Z)$ remains I for $\varepsilon \rightarrow 0$. Differentiating we obtain

$$T_Y \mathcal{V} = \{Z \in \mathbb{R}^{n \times k} \mid Z^T Y + Y^T Z = 0\}, \quad (9.2)$$

i.e., $Y^T Z$ is skew-symmetric. This represents $\frac{1}{2}k(k+1)$ conditions, thus $T_Y \mathcal{V}$ is of dimension $nk - \frac{1}{2}k(k+1)$.

For defining the normal space, we use the standard Euclidean inner product on $\mathbb{R}^{n \times k}$, i.e.,

$$\langle A, B \rangle = \text{trace}(A^T B) = \sum_{ij} a_{ij} b_{ij}, \quad (9.3)$$

whose corresponding norm is the Frobenius norm

$$\|A\|_F = \sqrt{\sum_{ij} a_{ij}^2}. \quad (9.4)$$

Then the normal space at Y is given by

$$N_Y \mathcal{V} = \{K \in \mathbb{R}^{n \times k} \mid K \perp T_Y \mathcal{V}\} = \{YS \mid S \text{ symmetric } k \times k \text{ matrix}\}. \quad (9.5)$$

To show this, we observe that the orthogonality $YS \perp T_Y \mathcal{V}$ follows from $\langle YS, Z \rangle = \text{trace}(SY^T Z) = \langle S, Y^T Z \rangle$ and the fact that any symmetric matrix A is orthogonal to any skew-symmetric matrix B .⁷ A dimension count (the matrix S has $\frac{1}{2}k(k+1)$ free elements) now shows us that the space defined in (9.5) fills the entire orthogonal complement of $T_Y \mathcal{V}$.

Orthogonality-Preserving Runge–Kutta Methods. Suppose now that we have to solve a differential equation $\dot{Y} = F(Y)$ on a Stiefel manifold \mathcal{V} . The orthogonality constraints $Y^T Y = I$ are preserved, if the derivative $F(Y)$ lies in the tangent space $T_Y \mathcal{V}$, i.e., if $F(Y)^T Y + Y^T F(Y) = 0$, for every $Y \in \mathcal{V}$ (weak invariants, see Sect. IV.4). In the (exceptional) case where they are in fact true invariants, i.e., if $F(Y)^T Y + Y^T F(Y) = 0$ for *all* $Y \in \mathbb{R}^{n \times k}$, then the orthogonality constraints are quadratic, and are therefore preserved exactly by the implicit Runge–Kutta methods of Sect. IV.2.1, in particular the Gauss methods. These methods give numerical solutions on the Stiefel manifold, but use function evaluations outside the manifold.

In the general case of only weak invariants, a standard approach for enforcing orthogonality is the introduction of *Lagrange multipliers*, which can be interpreted as artificial forces in the direction of the normal space keeping the solutions on the manifold. Due to the structure of $N_Y \mathcal{V}$ (see (9.5)), the problem becomes here

$$\dot{Y} = F(Y) + Y\Lambda, \quad Y^T Y = I \quad (9.6)$$

with a symmetric Lagrange multiplier matrix $\Lambda \in \mathbb{R}^{k \times k}$; see also Exercise 10. Any numerical method for differential-algebraic equations can now be applied, e.g.,

⁷ Indeed, split the sum in (9.3) in two parts $i < j$ and $i > j$, and interchange $i \leftrightarrow j$ in the second sum. Then both sums are identical with opposite sign.

appropriate Runge-Kutta methods as in Chap. VI and Sect. VII.4 of Hairer & Wanner (1996). A symmetric adaptation of Gauss methods to such problems is given by Jay (2005).

Below we shall study in great detail mechanical systems with constraints (see Sect. VII.1). In the case of orthogonality constraints, such problems can be treated successfully with Lobatto IIIA-IIIB partitioned Runge-Kutta methods, which in addition to orthogonality preserve other important geometric properties such as reversibility and symplecticity.

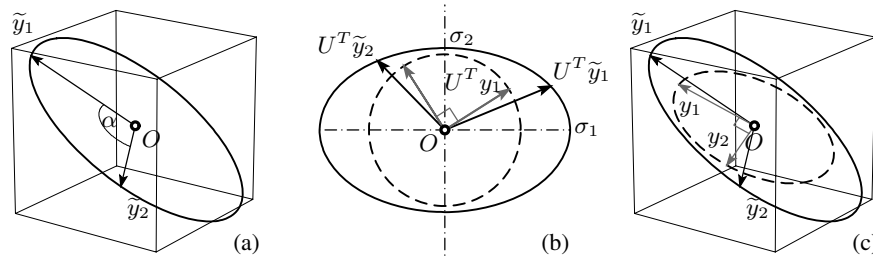


Fig. 9.1. Projection onto the Stiefel manifold using the singular value decomposition

Projection Methods. If we want to use the projection method of Algorithm 4.2, we have to perform, after every integration step, the projection (4.4), which requires to find for any given matrix \tilde{Y} a matrix $Y \in \mathcal{V}$ with

$$\|Y - \tilde{Y}\|_F \rightarrow \min. \quad (9.7)$$

This projection can be obtained as follows: if \tilde{Y} is not in \mathcal{V} (but close), then its column vectors $\tilde{y}_1, \dots, \tilde{y}_k$ will have norms different from 1 and/or their angles will not be right angles. These quantities determine an ellipsoid, if we require that these vectors represent conjugate diameters⁸ (see Fig. 9.1 (a)). This ellipsoid is then transformed to principal axes in \mathbb{R}^k by an orthogonal map U^T (picture (b)). We let $\sigma_1, \dots, \sigma_k$ be the length of these axes. If the coordinates are now divided by σ_i , then the ellipsoid becomes the unit sphere and the vectors $U^T \tilde{y}_i$ become orthonormal vectors $U^T y_i$. These vectors, when transformed back with U , lie in \mathcal{V} and are the projection we were searching for (picture (c)). For a proof of the optimality, see Exercise 21.

Connection with the Singular Value Decomposition. We have by construction that $U^T y_i = \Sigma^{-1} U^T \tilde{y}_i$ where $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_k)$. If we finally map these vectors by an orthogonal matrix V to the unit base, we see that $V \Sigma^{-1} U^T \tilde{Y} = I$, or

$$\tilde{Y} = U \Sigma V^T \quad (9.8)$$

which is the *singular value decomposition* of \tilde{Y} . This connection allows us to use standard software for our calculations. The projected matrix is then $Y = UV^T$.

⁸ Here we touch another of Stiefel's great ideas, the CG algorithm.

Remark 1. When the differential equation possesses some symmetry (see the next chapter), then the *symmetric* projection algorithm V.4.1 is preferable to be used instead.

Remark 2. The above procedure is equivalent to the one proposed by D. Higham (1997): the orthogonal projection is the first factor of the *polar decomposition* $\tilde{Y} = YR$ (where Y has orthonormal columns and R is symmetric positive definite). The equivalence is seen from the polar decomposition $\tilde{Y} = (UV^T)(V\Sigma V^T)$. A related procedure, where the first factor of the *QR decomposition* of \tilde{Y} is used instead of that of the polar decomposition, is proposed in Dieci, Russell & van Vleck (1994).

Tangent Space Parametrization. For the application of the methods of Sect. IV.5, in particular Subsection IV.5.3, to the case of Stiefel manifolds, we have to find the formulas for the projection (5.8) (see the wrap figure).

For a fixed Y , let $Y+Z$ be an arbitrary matrix in $Y + T_Y\mathcal{V}$, for which we search the projection $\psi_Y(Z)$ to \mathcal{V} . Because of the structure of $N_Y\mathcal{V}$ (see (9.5)), we have that

$$\psi_Y(Z) = Y + Z + YS \quad (9.9)$$

is a local parametrization of \mathcal{V} , if S is symmetric and if $\psi_Y(Z)^T \psi_Y(Z) = I$. This condition, when multiplied out, shows that S has to be a solution of the algebraic Riccati equation

$$S^2 + 2S + SY^T Z + Z^T Y S + Z^T Z = 0. \quad (9.10)$$

Observe that for $k = 1$, where the Stiefel manifold reduces to the unit sphere in \mathbb{R}^n , the equation (9.10) is a scalar quadratic equation and can be easily solved. For $k > 1$, it can be solved iteratively using the scheme (e.g., starting with $S_0 = 0$)

$$(I + Z^T Y)S_n + S_n(I + Y^T Z) = -Z^T Z - S_{n-1}^2.$$

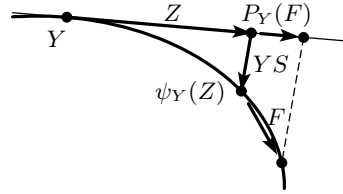
Using a Schur decomposition $Y^T Z = Q^T R Q$ (where Q is orthogonal and R upper triangular), the elements of $Q S_n Q^T$ can be computed successively starting from the left upper corner. We refer to the monograph of Mehrmann (1991) for a detailed discussion of the solution of linear and algebraic Riccati equations.

Next, we compute for the matrix F its orthogonal projection $P_Y(F)$ to $T_Y\mathcal{V}$, i.e., by (9.5), we have to find a symmetric matrix \tilde{S} such that $P_Y(F) = F - Y\tilde{S}$. The tangent condition $P_Y(F)^T Y + Y^T P_Y(F) = 0$ leads to $\tilde{S} = (F^T Y + Y^T F)/2$, so that

$$P_Y(F) = F - \frac{1}{2}(Y F^T Y + Y Y^T F). \quad (9.11)$$

With the parametrization $\psi_Y(Z)$ of (9.9) the transformed differential equation, when projected to the tangent space, yields

$$\dot{Z} = P_Y F(\psi_Y(Z)), \quad (9.12)$$



in complete analogy to (5.9). The numerical solution of (9.12) requires, for every function evaluation, the solution of the Riccati equation (9.10) and the computation of a projection onto the tangent space, each needing $\mathcal{O}(nk^2)$ operations. Compared with the projection method, the overhead (i.e., the computation apart from the evaluation of $F(Y)$) is more expensive, but the approach described here has the advantage that all evaluations of F are exactly on the manifold \mathcal{V} .

IV.9.2 Differential Equations on the Grassmann Manifold

The Grassmann manifold is obtained from the Stiefel manifold by identifying matrices in $\mathcal{V}_{n,k}$ that span the same subspace (see Fig. 9.2 (a)). Since any two such matrices result from each other by right multiplication with an orthogonal $k \times k$ matrix, the resulting manifold is the quotient manifold

$$\mathcal{G}_{n,k} = \mathcal{V}_{n,k}/\mathcal{O}(k). \quad (9.13)$$

An equivalence class $[Y] \in \mathcal{G}_{n,k}$ defines an orthogonal projection $P = YY^T$ of rank k , and conversely, every orthogonal basis of the range of P yields a representative $Y \in \mathcal{V}_{n,k}$. We can thus view the Grassmann manifold as

$$\mathcal{G}_{n,k} = \left\{ P \mid \begin{array}{l} P \text{ is an orthogonal projection onto} \\ \text{a } k\text{-dimensional subspace of } \mathbb{R}^n \end{array} \right\}. \quad (9.14)$$

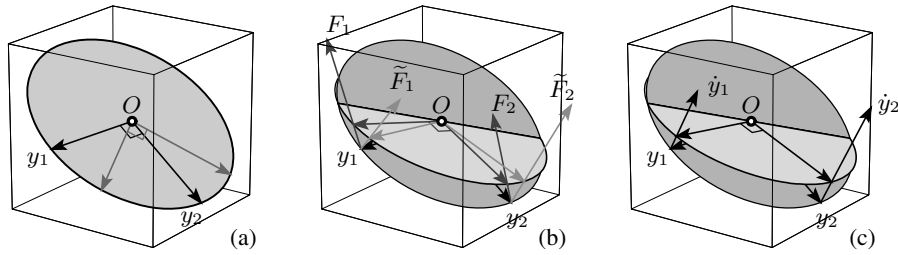


Fig. 9.2. Integration of a differential equation on the Grassmann manifold

The Tangent Space. The map $Y \mapsto P = YY^T$ from $\mathcal{V} \rightarrow \mathcal{G}$ has the tangent map (derivative)⁹

$$T_Y \mathcal{V} \rightarrow T_P \mathcal{G} : \quad \delta Y \mapsto \delta P = \delta Y Y^T + Y \delta Y^T, \quad (9.15)$$

and we wish to apply all the methods for $T_Y \mathcal{V}$ from the arsenal of the preceding section to problems in $T_P \mathcal{G}$. However, the dimension of $T_P \mathcal{G}$ is by $\frac{1}{2}k(k-1)$ lower than the dimension of $T_Y \mathcal{V}$. This difference is the dimension of $\mathcal{O}(k)$ and also of

⁹ Here we write δY for tangent matrices at Y (what has been Z in (9.2)), and similarly for other matrices; Lagrange's δ -notation here becomes preferable, since we will have, especially in the next subsection, more and more matrices moving around.

$\mathfrak{so}(k)$, the vector space of skew-symmetric $k \times k$ matrices. The key idea is now the following: if we replace the condition from (9.2), $Y^T \delta Y$ skew-symmetric, by $Y^T \delta Y = 0$, then we remove precisely the superfluous degrees of freedom. Indeed, the extended tangent map

$$T_Y \mathcal{V} \rightarrow T_P \mathcal{G} \times \mathfrak{so}(k) : \delta Y \mapsto (\delta Y Y^T + Y \delta Y^T, Y^T \delta Y) \quad (9.16)$$

is an isomorphism, since it is readily seen to have zero null-space and the dimensions of the vector spaces agree. The tangent space is thus characterized as

$$T_P \mathcal{G} = \{\delta P = \delta Y Y^T + Y \delta Y^T \mid Y^T \delta Y = 0\}, \quad (9.17)$$

and every $\delta P \in T_P \mathcal{G}$ corresponds to a unique δY with $Y^T \delta Y = 0$. Note that this condition on δY does not depend on the representative Y of $[Y]$.

Differential Equations. Consider now a differential equation on \mathcal{G} ,

$$\dot{P} = G(P), \quad (9.18)$$

with a vector field G on \mathcal{G} . The condition $G(P) \in T_P \mathcal{G}$ means, since the tangent map (9.15) is onto, that there exists for $P = Y Y^T$ a vector $F(Y)$ such that

$$G(P) = F(Y) Y^T + Y F(Y)^T \quad \text{with} \quad F^T Y + Y^T F = 0 \quad (9.19)$$

i.e., $F(Y) \in T_Y \mathcal{V}$. However, from a given initial position Y , there are many F which produce the same movement G of the subspace represented by P (see Fig. 9.2 (b)). By (9.16), the movement of Y becomes unique if we require that this movement is *orthogonal* to the subspace (see Fig. 9.2 (c)),

$$Y^T \dot{Y} = 0. \quad (9.20)$$

Multiplying the derivative $\dot{P} = \dot{Y} Y^T + Y \dot{Y}^T$ with Y^T from the left, we obtain, under condition (9.20), $Y^T \dot{P} = \dot{Y}^T$ and, by (9.18) and (9.19), $\dot{Y} = Y F^T Y + F$ or

$$\dot{Y} = (I - Y Y^T) F(Y). \quad (9.21)$$

Geometrically, this means that the vector $F(Y)$, which could be chosen arbitrarily in $T_Y \mathcal{V}$, is projected to the orthogonal complement of the subspace spanned by Y or $P = Y Y^T$. The derivative \dot{Y} in (9.21) is independent of the particular choice of F .

Equation (9.21) is a differential equation on the Stiefel manifold \mathcal{V} that can be solved numerically by the methods described in the previous subsection.

Example 9.1 (Oja Flow). A basic example arises in neural networks (Oja 1989): solutions on $\mathcal{V}_{n,k}$ of the differential equation

$$\dot{Y} = (I - Y Y^T) A Y \quad (9.22)$$

with a constant symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$ tend to an orthogonal basis of an invariant subspace of A as $t \rightarrow \infty$ (Yan, Helmke & Moore 1994).

A naïve comparison of this equation with (9.21) would lead to $F(Y) = AY$, but this function does not satisfy the tangent condition $F^T Y + Y^T F = 0$ from (9.19). So we use the fact that $(I - YY^T)^2 = I - YY^T$ and set $F(Y) = (I - YY^T)AY$. With this, $G(P)$ from (9.18) and (9.19) becomes

$$\dot{P} = (I - P)AP + PA(I - P). \quad (9.23)$$

We have obtained the result that equation (9.22) can be viewed as a differential equation on the Grassmann manifold $\mathcal{G}_{n,k}$.

However, for the numerical integration it is more practical to work with (9.22).

IV.9.3 Dynamical Low-Rank Approximation

Low-rank approximation of large matrices is a basic model reduction technique in many application areas, such as image compression and latent semantic indexing in information retrieval; see for example Simon & Zha (2000). Here, we consider the task of computing low rank approximations to matrices $A(t) \in \mathbb{R}^{m \times n}$ depending smoothly on t . At any time t , a best approximation to $A(t)$ of rank k is a matrix $X(t)$ in the manifold $\mathcal{M}_k = \mathcal{M}_k^{m \times n}$ of rank- k matrices that satisfies

$$X(t) \in \mathcal{M}_k \quad \text{such that} \quad \|X(t) - A(t)\|_F = \min! \quad (9.24)$$

The problem is solved by a singular value decomposition of $A(t)$, truncating all singular values after the r largest ones. When the matrix is so large that a complete singular value decomposition is not feasible, a standard approach to obtain an approximate solution is based on the Lanczos bidiagonalization process with $A(t)$, as discussed in Simon & Zha (2000).

Following Koch & Lubich (2005), we here consider instead the low-rank approximation $Y(t) \in \mathcal{M}_k$ determined from the condition that for every t the derivative $\dot{Y}(t)$, which is in the tangent space $T_{Y(t)}\mathcal{M}_k$, be chosen as

$$\dot{Y}(t) \in T_{Y(t)}\mathcal{M}_k \quad \text{such that} \quad \|\dot{Y}(t) - \dot{A}(t)\|_F = \min! \quad (9.25)$$

This is complemented with an initial condition, ideally $Y(t_0) = X(t_0)$. For given $Y(t)$, the derivative $\dot{Y}(t)$ is obtained by a *linear* projection, though onto a solution-dependent vector space. Problem (9.25) yields a differential equation on \mathcal{M}_k . We will see that with a suitable factorization of rank- k matrices, we obtain a system of differential equations for the factors that is well-suited for numerical integration. The differential equations contain only the increments $\dot{A}(t)$, which may be much sparser than the full data matrix $A(t)$.

Koch & Lubich (2005) show that $Y(t)$ yields a quasi-optimal approximation on intervals where a good smooth approximation exists. It must be noted, however, that the best rank- k approximation $X(t)$ may have discontinuities, which cannot be captured in $Y(t)$. This is already seen from the example of finding a rank-1 approximation to $\text{diag}(e^{-t}, e^t)$, where starting from $t_0 < 0$ yields $X(t) = Y(t) = \text{diag}(e^{-t}, 0)$ for $t < 0$, but $Y(t) = \text{diag}(e^{-t}, 0)$ and $X(t) = \text{diag}(0, e^t)$ for $t > 0$.

The best approximation $X(t)$ has a discontinuity at $t = 0$, caused by a crossing of singular values of which one is inside and the other one outside the approximation. An algorithmic remedy is to restart (9.25) at regular intervals.

In contrast to (9.24), the approach (9.25) extends immediately to the low-rank approximation of solutions of matrix differential equations $\dot{A} = F(A)$. Here, $\dot{A}(t)$ in (9.25) is simply replaced by the approximation $F(Y(t))$, which yields the minimum-defect low-rank approximation $Y(t)$ by choosing

$$\dot{Y} \in T_Y \mathcal{M}_k \quad \text{such that} \quad \|\dot{Y} - F(Y)\|_F = \min! \quad (9.26)$$

An approach of this type is of common use in quantum dynamics, where the physical model reduction of the multivariate Schrödinger equation by the analogue of (9.26) is known as the Dirac-Frenkel time-dependent variational principle, after Dirac (1930) and Frenkel (1934); see also Beck, Jäckle, Worth & Meyer (2000) and Sect. VII.6.

Decompositions of Rank- k Matrices and of Their Tangent Matrices. Every real rank- k matrix of dimension $m \times n$ can be written in the form

$$Y = USV^T \quad (9.27)$$

where $U \in \mathcal{V}_{m,k}$ and $V \in \mathcal{V}_{n,k}$ have orthonormal columns, and $S \in \mathbb{R}^{k \times k}$ is nonsingular. The singular value decomposition yields S diagonal, but here we do not assume a special form of S . The representation (9.27) is not unique: replacing U by $\tilde{U} = UP$ and V by $\tilde{V} = VQ$ with orthogonal matrices $P, Q \in \mathcal{O}(k)$ and correspondingly S by $\tilde{S} = P^T S Q$, yields the same matrix $Y = USV^T = \tilde{U} \tilde{S} \tilde{V}^T$.

As a substitute for the non-uniqueness in (9.27), we use – as in the previous subsection – a unique decomposition in the tangent space. Every tangent matrix $\delta Y \in T_Y \mathcal{M}_k$ at $Y = USV^T$ is of the form (see Exercise 23)

$$\delta Y = \delta U S V^T + U \delta S V^T + U S \delta V^T, \quad (9.28)$$

where $\delta S \in \mathbb{R}^{k \times k}$ and $\delta U \in T_U \mathcal{V}_{m,k}$, $\delta V \in T_V \mathcal{V}_{n,k}$. Conversely, $\delta S, \delta U, \delta V$ are uniquely determined by δY if we impose the orthogonality constraints

$$U^T \delta U = 0, \quad V^T \delta V = 0. \quad (9.29)$$

Equations (9.28) and (9.29) yield

$$\begin{aligned} \delta S &= U^T \delta Y V, \\ \delta U &= (I - UU^T) \delta Y V S^{-1}, \\ \delta V &= (I - VV^T) \delta Y^T U S^{-T}. \end{aligned} \quad (9.30)$$

Formulas (9.28) and (9.30) establish an isomorphism between the subspace

$$\{(\delta S, \delta U, \delta V) \in \mathbb{R}^{k \times k} \times \mathbb{R}^{m \times k} \times \mathbb{R}^{n \times k} \mid U^T \delta U = 0, V^T \delta V = 0\}$$

and the tangent space $T_Y \mathcal{M}_k$.

Differential Equations for the Factors. The minimization condition (9.25) is equivalent to the orthogonal projection of $\dot{A}(t)$ onto the tangent space $T_{Y(t)}\mathcal{M}_k$: find $\dot{Y} \in T_Y\mathcal{M}_k$ (we omit the argument t) satisfying

$$\langle \dot{Y} - \dot{A}, \delta Y \rangle = 0 \quad \text{for all } \delta Y \in T_Y\mathcal{M}_k, \quad (9.31)$$

with the Frobenius inner product $\langle A, B \rangle = \text{trace}(A^T B)$. With this formulation we derive differential equations for the factors in the representation (9.27).

Theorem 9.2. For $Y = USV^T \in \mathcal{M}_k$ with nonsingular $S \in \mathbb{R}^{k \times k}$ and with $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$ having orthonormal columns, condition (9.25) or (9.31) is equivalent to $\dot{Y} = \dot{U}SV^T + U\dot{S}V^T + US\dot{V}^T$, where

$$\begin{aligned} \dot{S} &= U^T \dot{A} V \\ \dot{U} &= (I - UU^T) \dot{A} V S^{-1} \\ \dot{V} &= (I - VV^T) \dot{A}^T U S^{-T}. \end{aligned} \quad (9.32)$$

Proof. For $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ and $B \in \mathbb{R}^{m \times n}$, we use the identity

$$\langle uv^T, B \rangle = u^T B v.$$

In view of (9.29) we require $U^T \dot{U} = V^T \dot{V} = 0$ along the solution trajectory in order to define a unique representation of \dot{Y} . We first substitute $\delta Y = u_i v_j^T$, for $i, j = 1, \dots, k$, in (9.31), where u_i, v_j denote the columns of U, V , respectively. This is of the form (9.27) with $\delta U = \delta V = 0$ and one non-zero element in δS . In this way we find $\dot{S} = U^T \dot{A} V$. Similarly, choosing $\delta Y = \sum_{j=1}^k \delta u s_{ij} v_j^T$, $i = 1, \dots, k$, where $\delta u \in \mathbb{R}^m$ is arbitrary with $U^T \delta u = 0$, we obtain the stated differential equation for U , and likewise for $\delta Y = \sum_{j=1}^k u_j s_{ji} \delta v^T$ with $V^T \delta v = 0$ the differential equation for V . \square

The differential equations (9.32) are closely related to differential equations for other smooth matrix decompositions, in particular the smooth singular value decomposition; see, e.g., Dieci & Eirola (1999) and Wright (1992). Unlike the differential equations for singular values given there, the equations (9.32) have no singularities at points where singular values of $Y(t)$ coalesce.

For the minimum-defect low-rank approximation (9.26) of a matrix differential equation $\dot{A} = F(A)$, we just need to replace \dot{A} by $F(Y)$ for $Y = USV^T$ in the differential equations (9.32).

The matrices $U(t)$ and $V(t)$ evolve on Stiefel manifolds. The differential equations (9.32) can thus be solved numerically by the methods discussed in Sect. IV.9.1.

IV.10 Exercises

1. Prove that the symplectic Euler method (I.1.9) conserves quadratic invariants of the form (2.5). Explain the “0” entries of Table (I.2.1).

2. Prove that under condition (2.3) a Runge–Kutta method preserves all invariants of the form $I(y) = y^T C y + d^T y + c$.
3. Prove that an s -stage diagonally implicit Runge–Kutta method (i.e., $a_{ij} = 0$ for $i < j$) satisfies the condition (2.3) if and only if it is equivalent to a composition $\Phi_{b_s h} \circ \dots \circ \Phi_{b_1 h}$ based on the implicit midpoint rule.
4. Prove the following statements: a) If a partitioned Runge–Kutta method conserves general quadratic invariants $p^T C p + 2p^T D q + q^T E q$, then each of the two Runge–Kutta methods has to conserve quadratic invariants separately.
b) If both methods, $\{b_i, a_{ij}\}$ and $\{\widehat{b}_i, \widehat{a}_{ij}\}$ are irreducible, satisfy (2.3) and if (2.7)–(2.8) hold, then we have $b_i = \widehat{b}_i$ and $a_{ij} = \widehat{a}_{ij}$ for all i, j .
5. Prove that the Gauss methods are the only collocation methods satisfying (2.3). *Hint.* Use the ideas of the proof of Lemma 13.9 in Hairer & Wanner (1996).
6. Discontinuous collocation methods with either $b_1 \neq 0$ or $b_s \neq 0$ (Definition II.1.7) cannot satisfy the criterion (2.3).
7. (Sanz-Serna & Abia 1991, Saito, Sugiura & Mitsui 1992). The condition (2.3) acts as simplifying assumption for the order conditions of Runge–Kutta methods. Assume that the order conditions are satisfied for the trees u and v . Prove that it is satisfied for $u \circ v$ if and only if it is satisfied for $v \circ u$, and that it is automatically satisfied for trees of the form $u \circ u$.
Remark. $u \circ v$ denotes the Butcher product introduced in Sect. VI.7.2.
8. If L_0 is a symmetric, tridiagonal matrix that is sufficiently close to $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, where $\lambda_1 > \lambda_2 > \dots > \lambda_n$ are the eigenvalues of L_0 , then the solution of (3.5) with $B(L) = L_+ - L_+^T$ converges exponentially fast to the diagonal matrix Λ . Hence, the numerical solution of (3.5) gives an algorithm for the computation of the eigenvalues of the matrix L_0 .
Hint. Let β_1, \dots, β_n be the entries in the diagonal of L , and $\alpha_1, \dots, \alpha_{n-1}$ those in the subdiagonal. Assume that $|\beta_k(0) - \lambda_k| \leq R/3$ and $|\alpha_k(0)| \leq R$ with some sufficiently small R . Prove that $\beta_k(t) - \beta_{k+1}(t) \geq \mu - R$ and $|\alpha_k(t)| \leq R e^{-(\mu-R)t}$ for all $t \geq 0$, where $\mu = \min_k(\lambda_k - \lambda_{k+1}) > 0$.
9. Elaborate Example 4.5 for the special case where Y is a matrix of dimension 2. In particular, show that (4.6) is the same as (4.5), and check the formulas for the simplified Newton iterations.
10. (Brenan, Campbell & Petzold (1996), Sect. 2.5.3). Consider the differential equation $\dot{y} = f(y)$ with known invariants $g(y) = \text{Const}$, and assume that $g'(y)$ has full rank. Prove by differentiation of the constraints that, for initial values satisfying $g(y_0) = 0$, the solution of the differential-algebraic equation (DAE)

$$\begin{aligned} \dot{y} &= f(y) + g'(y)^T \mu \\ 0 &= g(y) \end{aligned}$$

also solves the differential equation $\dot{y} = f(y)$.

Remark. Most methods for DAEs (e.g., stiffly accurate Runge–Kutta methods or BDF methods) lead to numerical integrators that preserve exactly the constraints $g(y) = 0$. The difference from the projection method of Sect. IV.4 is that here the internal stages also satisfy the constraint.

11. Prove that $\mathrm{SL}(n)$ is a Lie group of dimension $n^2 - 1$, and that $\mathfrak{sl}(n)$ is its Lie algebra (see Table 6.1 for the definitions of $\mathrm{SL}(n)$ and $\mathfrak{sl}(n)$).
12. Let G be a matrix Lie group and \mathfrak{g} its Lie algebra. Prove that for $Y \in G$ and $A \in \mathfrak{g}$ we have $YAY^{-1} \in \mathfrak{g}$.
Hint. Consider the path $\gamma(t) = Y\alpha(t)Y^{-1}$.
13. Consider a problem $\dot{Y} = A(Y)Y$, for which $A(Y) \in \mathfrak{so}(n)$ whenever $Y \in \mathrm{O}(n)$, but where $A(Y)$ is an arbitrary matrix for $Y \notin \mathrm{O}(n)$.
 a) Prove that $Y_0 \in \mathrm{O}(n)$ implies $Y(t) \in \mathrm{O}(n)$ for all t .
 b) Show by a counter-example that the numerical solution of the implicit midpoint rule does not necessarily stay in $\mathrm{O}(n)$.
14. (Feng Kang & Shang Zai-jiu 1995). Let $R(z) = (1 + z/2)/(1 - z/2)$ be the stability function of the implicit midpoint rule. Prove that for $A \in \mathfrak{sl}(3)$ we have

$$\det R(hA) = 1 \quad \Leftrightarrow \quad \det A = 0.$$

15. (Iserles & Nørsett 1999). Introducing $y_1 = y$ and $y_2 = \dot{y}$, write the problem

$$\ddot{y} + ty = 0, \quad y(0) = 1, \quad \dot{y}(0) = 0$$

in the form (7.6). Then apply the numerical method of Example 7.4 with different step sizes on the interval $0 \leq t \leq 100$. Compare the result with that obtained by fourth order classical (explicit or implicit) Runge–Kutta methods.
Remark. If $A(t)$ in (7.6) (or $A(t, y)$ in (8.1)) are much smoother than the solution $y(t)$, then Lie group methods are usually superior to standard integrators, because Lie group methods approximate $A(t)$, whereas standard methods approximate the solution $y(t)$ by polynomials.

16. Deduce the BCH formula from the Magnus expansion (IV.7.5).
Hint. For constant matrices A and B consider the matrix function $A(t)$ defined by $A(t) = B$ for $0 \leq t \leq 1$ and $A(t) = A$ for $1 \leq t \leq 2$.
17. (Rodrigues formula, see Marsden & Ratiu (1999), page 291). Prove that

$$\exp(\Omega) = I + \frac{\sin \alpha}{\alpha} \Omega + \frac{1}{2} \left(\frac{\sin(\alpha/2)}{\alpha/2} \right)^2 \Omega^2 \quad \text{for} \quad \Omega = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}$$

where $\alpha = \sqrt{\omega_1^2 + \omega_2^2 + \omega_3^2}$. This formula allows for an efficient implementation of the Lie group methods in $\mathrm{O}(3)$.

18. The solution of $\dot{Y} = A(Y)Y$, $Y(0) = Y_0$, is given by $Y(t) = \exp(\Omega(t))Y_0$, where $\Omega(t)$ solves the differential equation (8.9). Compute the first terms of the t -expansion of $\Omega(t)$.
Result. $\Omega(t) = tA(Y_0) + \frac{t^2}{2}A'(Y_0)A(Y_0)Y_0 + \frac{t^3}{6}(A'(Y_0)^2A(Y_0)Y_0^2 + A'(Y_0)A(Y_0)^2Y_0 + A''(Y_0)(A(Y_0)Y_0, A(Y_0)Y_0) - \frac{1}{2}[A(Y_0), A'(Y_0)A(Y_0)Y_0])$.
19. Consider the 2-stage Gauss method of order $p = 4$. In the corresponding Lie group method, eliminate the presence of Ω in $[\Omega, A]$ by iteration, and neglect higher order commutators. Show that this leads to

$$\begin{aligned}\Omega_1 &= h\left(\frac{1}{4}A_1 + \left(\frac{1}{4} - \frac{\sqrt{3}}{6}\right)A_2\right) - \frac{h^2}{2}\left(-\frac{1}{12} + \frac{\sqrt{3}}{24}\right)[A_1, A_2] \\ \Omega_2 &= h\left(\left(\frac{1}{4} + \frac{\sqrt{3}}{6}\right)A_1 + \frac{1}{4}A_2\right) - \frac{h^2}{2}\left(\frac{1}{12} + \frac{\sqrt{3}}{24}\right)[A_1, A_2] \\ y_1 &= \exp\left(h\left(\frac{1}{2}A_1 + \frac{1}{2}A_2\right) - h^2\frac{\sqrt{3}}{12}[A_1, A_2]\right)y_0,\end{aligned}$$

where $A_i = A(Y_i)$ and $Y_i = \exp(\Omega_i)y_0$. Prove that this is a Lie group method of order 4. Is it symmetric?

20. In Zanna (1999) a Lie group method similar to that of Exercise 19 is presented. The only difference is that the coefficients $(-1/12 + \sqrt{3}/24)$ and $(1/12 + \sqrt{3}/24)$ in the formulas for Ω_1 and Ω_2 are replaced with $(-5/72 + \sqrt{3}/24)$ and $(5/72 + \sqrt{3}/24)$, respectively. Is there an error somewhere? Are both methods of order 4?
21. Show that for given \tilde{Y} the solution of problem (9.7) is $Y = UV^T$, where $\tilde{Y} = U\Sigma V^T$ is the singular value decomposition of \tilde{Y} .
Hint. Since $\|USV^T\|_F = \|S\|_F$ holds for all orthogonal matrices U and V , it is sufficient to consider the case $\tilde{Y} = (\Sigma, 0)^T$ with $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_k)$. Prove that $\|(\Sigma, 0)^T - Y\|_F^2 \geq \sum_{i=1}^k (\sigma_i - 1)^2$ for all matrices Y satisfying $Y^TY = I$.
22. Show that the solution of the matrix differential equation $\dot{Y} = A(t)Y$ on $\mathbb{R}^{n \times k}$, with initial values $Y_0 \in \mathcal{V}_{n,k}$, can be decomposed as

$$Y(t) = U(t)S(t), \quad \text{where } U(t) \in \mathcal{V}_{n,k}, S(t) \in \mathbb{R}^{k \times k}$$

satisfy the differential equations

$$\dot{S} = U^T A U S, \quad \dot{U} = (I - UU^T)AU$$

with initial values $S_0 = I, U_0 = Y_0$.

Remark: These differential equations can be used for the computation of Lyapunov exponents as an alternative to the differential equations discussed in Bridges & Reich (2001) and Dieci, Russell & van Vleck (1997).

23. Consider the map $\text{GL}(k) \times \mathcal{V}_{m,k} \times \mathcal{V}_{n,k} \rightarrow \mathcal{M}_k$ that associates to (S, U, V) the rank- k matrix $Y = USV^T$. Show that the extended tangent map

$$\begin{aligned}\mathbb{R}^{k \times k} \times T_U \mathcal{V}_{m,k} \times T_V \mathcal{V}_{n,k} &\rightarrow T_Y \mathcal{M}_k \times \mathfrak{so}(k) \times \mathfrak{so}(k) \\ (\delta S, \delta U, \delta V) &\mapsto (\delta USV^T + U\delta SV^T + US\delta V^T, U^T \delta U, V^T \delta V)\end{aligned}$$

is an isomorphism.

24. Let $A(t) \in \mathbb{R}^{n \times n}$ be symmetric and depend smoothly on t . Show that the solution $P(t) \in \mathcal{G}_{n,k}$ of the dynamical low-rank approximation problem on the Grassmann manifold,

$$\dot{P} \in T_P \mathcal{G}_{n,k} \quad \text{with} \quad \|\dot{P} - \dot{A}\|_F = \min!,$$

is given as $P = YY^T$ where $Y \in \mathcal{V}_{n,k}$ solves the differential equation

$$\dot{Y} = (I - YY^T)\dot{A}Y.$$

Chapter V.

Symmetric Integration and Reversibility

Symmetric methods of this chapter and symplectic methods of the next chapter play a central role in the geometric integration of differential equations. We discuss reversible differential equations and reversible maps, and we explain how symmetric integrators are related to them. We study symmetric Runge–Kutta and composition methods, and we show how standard approaches for solving differential equations on manifolds can be symmetrized. A theoretical explanation of the excellent long-time behaviour of symmetric methods applied to reversible differential equations will be given in Chap. XI.

V.1 Reversible Differential Equations and Maps

Conservative mechanical systems have the property that inverting the initial direction of the velocity vector and keeping the initial position does not change the solution trajectory, it only inverts the direction of motion. Such systems are “reversible”. We extend this notion to more general situations.

Definition 1.1. Let ρ be an invertible linear transformation in the phase space of $\dot{y} = f(y)$. This differential equation and the vector field $f(y)$ are called ρ -reversible if

$$\rho f(y) = -f(\rho y) \quad \text{for all } y. \quad (1.1)$$

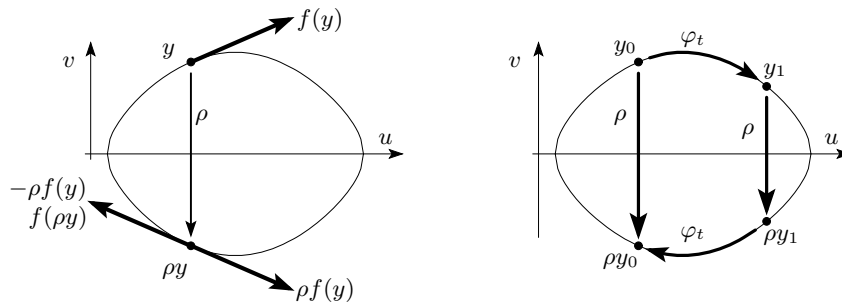


Fig. 1.1. Reversible vector field (left picture) and reversible map (right picture)

This property is illustrated in the left picture of Fig. 1.1. For ρ -reversible differential equations the exact flow $\varphi_t(y)$ satisfies

$$\rho \circ \varphi_t = \varphi_{-t} \circ \rho = \varphi_t^{-1} \circ \rho \quad (1.2)$$

(see the picture to the right in Fig. 1.1). The right identity is a consequence of the group property $\varphi_t \circ \varphi_s = \varphi_{t+s}$, and the left identity follows from

$$\begin{aligned} \frac{d}{dt}(\rho \circ \varphi_t)(y) &= \rho f(\varphi_t(y)) = -f((\rho \circ \varphi_t)(y)) \\ \frac{d}{dt}(\varphi_{-t} \circ \rho)(y) &= -f((\varphi_{-t} \circ \rho)(y)), \end{aligned}$$

because all expressions of (1.2) satisfy the same differential equation with the same initial value $(\rho \circ \varphi_0)(y) = (\varphi_0 \circ \rho)(y) = \rho y$. Formula (1.2) motivates the following definition.

Definition 1.2. A map $\Phi(y)$ is called ρ -reversible if

$$\rho \circ \Phi = \Phi^{-1} \circ \rho.$$

Example 1.3. An important example is the partitioned system

$$\dot{u} = f(u, v), \quad \dot{v} = g(u, v), \quad (1.3)$$

where $f(u, -v) = -f(u, v)$ and $g(u, -v) = g(u, v)$. Here, the transformation ρ is given by $\rho(u, v) = (u, -v)$. If we call a vector field or a map *reversible* (without specifying the transformation ρ), we mean that it is ρ -reversible with this particular ρ . All second order differential equations $\ddot{u} = g(u)$ written as $\dot{u} = v$, $\dot{v} = g(u)$ are reversible. As a first implication of reversibility on the dynamics we mention the following fact: if u and v are scalar, and if (1.3) is reversible, then any solution that crosses the u -axis twice is periodic (Exercise 5, see also the solution of the pendulum problem in Fig. I.1.4).

It is natural to search for numerical methods that produce a reversible numerical flow when they are applied to a reversible differential equation. We then expect the numerical solution to have long-time behaviour similar to that of the exact solution; see Chap. XI for more precise statements. It turns out that the ρ -reversibility of a numerical one-step method is closely related to the concept of symmetry.

Thus the method is theoretically *symmetrical* or *reversible*, a terminology we have never seen applied elsewhere.

(P.C. Hammer & J.W. Hollingsworth 1955)

Definition 1.4. A numerical one-step method Φ_h is called *symmetric* or *time-reversible*,¹ if it satisfies

$$\Phi_h \circ \Phi_{-h} = id \quad \text{or equivalently} \quad \Phi_h = \Phi_{-h}^{-1}.$$

¹ The study of symmetric methods has its origin in the development of extrapolation methods (Gragg 1965, Stetter 1973), because the global error admits an asymptotic expansion in even powers of h . The notion of time-reversible methods is more common in the Computational Physics literature (Buneman 1967).

With the Definition II.3.1 of the adjoint method (i.e., $\Phi_h^* = \Phi_{-h}^{-1}$), the condition for symmetry reads $\Phi_h = \Phi_h^*$. A method $y_1 = \Phi_h(y_0)$ is symmetric if exchanging $y_0 \leftrightarrow y_1$ and $h \leftrightarrow -h$ leaves the method unaltered. In Chap. I we have already encountered the implicit midpoint rule (I.1.7) and the Störmer–Verlet scheme (I.1.17), both of which are symmetric. Many more symmetric methods will be given in the following sections.

Theorem 1.5. *If a numerical method, applied to a ρ -reversible differential equation, satisfies*

$$\rho \circ \Phi_h = \Phi_{-h} \circ \rho, \quad (1.4)$$

then the numerical flow Φ_h is a ρ -reversible map if and only if Φ_h is a symmetric method.

Proof. As a consequence of (1.4) the numerical flow Φ_h is ρ -reversible if and only if $\Phi_{-h} \circ \rho = \Phi_h^{-1} \circ \rho$. Since ρ is an invertible transformation, this is equivalent to the symmetry of the method Φ_h . \square

Similarly, it is also true that a symmetric method is ρ -reversible if and only if the ρ -compatibility condition (1.4) holds.

Compared to the symmetry of the method, condition (1.4) is much less restrictive. It is automatically satisfied by most numerical methods. Let us briefly discuss the validity of (1.4) for different classes of methods.

- *Runge–Kutta methods* (explicit or implicit) satisfy (1.4) without any restriction other than (1.1) on the vector field (Stoffer 1988). Let us illustrate the proof with the explicit Euler method $\Phi_h(y_0) = y_0 + hf(y_0)$:

$$(\rho \circ \Phi_h)(y_0) = \rho y_0 + h\rho f(y_0) = \rho y_0 - hf(\rho y_0) = \Phi_{-h}(\rho y_0).$$

- *Partitioned Runge–Kutta methods* applied to a partitioned system (1.3) satisfy the condition (1.4) if $\rho(u, v) = (\rho_1(u), \rho_2(v))$ with invertible ρ_1 and ρ_2 . The proof is the same as for Runge–Kutta methods. Notice that the mapping $\rho(u, v) = (u, -v)$ of Example 1.3 is of this special form.
- *Composition methods.* If two methods Φ_h and Ψ_h satisfy (1.4), then so does the adjoint Φ_h^* and the composition $\Phi_h \circ \Psi_h$. Consequently, the composition methods (3.1) and (3.2) below, which compose a basic method Φ_h and its adjoint with different step sizes, have the property (1.4) provided the basic method Φ_h has it.
- *Splitting methods* are based on a splitting $\dot{y} = f^{[1]}(y) + f^{[2]}(y)$ of the differential equation. If both vector fields, $f^{[1]}(y)$ and $f^{[2]}(y)$, satisfy (1.1), then their exact flows $\varphi_h^{[1]}$ and $\varphi_h^{[2]}$ satisfy (1.2). In this situation, the splitting method (II.5.6) has the property (1.4).
- For *differential equations on manifolds* we have to assume that ρ maps \mathcal{M} to \mathcal{M} . Otherwise, condition (1.1) does not make sense. For the projection method of Algorithm IV.4.2 with orthogonal projection onto the manifold we have: if the basic method satisfies (1.4) and if ρ is an orthogonal matrix, then it satisfies (1.4) as well. This follows from the fact that the tangent and normal spaces satisfy

$T_{\rho y}\mathcal{M} = \rho T_y\mathcal{M}$ and $N_{\rho y}\mathcal{M} = \rho^{-T} N_y\mathcal{M}$, respectively. A similar result holds for methods based on local coordinates, if the local parametrization is well chosen. For example, this is the case if $\rho\psi(z)$ is the parametrization at ρy_0 whenever $\psi(z)$ is the parametrization at y_0 .

V.2 Symmetric Runge–Kutta Methods

We give a characterization of symmetric methods of Runge–Kutta type and mention some important examples.

V.2.1 Collocation and Runge–Kutta Methods

Symmetric collocation methods are characterized by the symmetry of the collocation points with respect to the midpoint of the integration step.

Theorem 2.1. *The adjoint method of a collocation method (Definition II.1.3) based on c_1, \dots, c_s is a collocation method based on c_1^*, \dots, c_s^* , where*

$$c_i^* = 1 - c_{s+1-i}. \quad (2.1)$$

In the case that $c_i = 1 - c_{s+1-i}$ for all i , the collocation method is symmetric.

The adjoint method of a discontinuous collocation method (Definition II.1.7) based on b_1, b_s and c_2, \dots, c_{s-1} is a discontinuous collocation method based on b_1^, b_s^* and c_2^*, \dots, c_{s-1}^* , where*

$$b_1^* = b_s, \quad b_s^* = b_1 \quad \text{and} \quad c_i^* = 1 - c_{s+1-i}. \quad (2.2)$$

In the case that $b_1 = b_s$ and $c_i = 1 - c_{s+1-i}$ for all i , the discontinuous collocation method is symmetric.

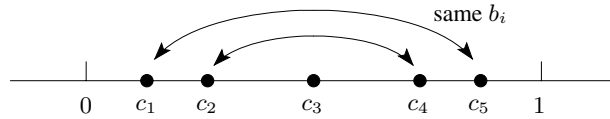


Fig. 2.1. Symmetry of collocation methods

Proof. Exchanging $(t_0, y_0) \leftrightarrow (t_1, y_1)$ and $h \leftrightarrow -h$ in the definition of a collocation method we get $u(t_1) = y_1$, $\dot{u}(t_1 - c_i h) = f(t_1 - c_i h, u(t_1 - c_i h))$, and $y_0 = u(t_1 - h)$. Inserting $t_1 = t_0 + h$ this yields the collocation method based on c_i^* of (2.1). Observe that the c_i^* can be arbitrarily permuted. For discontinuous collocation methods the proof is similar. \square

The preceding theorem immediately yields the following result.

Corollary 2.2. *The Gauss formulas (Table II.1.1), as well as the Lobatto IIIA (Table II.1.2) and Lobatto IIIB formulas (Table II.1.4) are symmetric integrators.* \square

Theorem 2.3 (Stetter 1973, Wanner 1973). *The adjoint method of an s -stage Runge–Kutta method (II.1.4) is again an s -stage Runge–Kutta method. Its coefficients are given by*

$$a_{ij}^* = b_{s+1-j} - a_{s+1-i, s+1-j}, \quad b_i^* = b_{s+1-i}. \quad (2.3)$$

If

$$a_{s+1-i, s+1-j} + a_{ij} = b_j \quad \text{for all } i, j, \quad (2.4)$$

then the Runge–Kutta method (II.1.4) is symmetric.²

Proof. Exchanging $y_0 \leftrightarrow y_1$ and $h \leftrightarrow -h$ in the Runge–Kutta formulas yields

$$k_i = f\left(y_0 + h \sum_{j=1}^s (b_j - a_{ij})k_j\right), \quad y_1 = y_0 + h \sum_{i=1}^s b_i k_i. \quad (2.5)$$

Since the values $\sum_{j=1}^s (b_j - a_{ij}) = 1 - c_i$ appear in reverse order, we replace k_i by k_{s+1-i} in (2.5), and then we substitute all indices i and j by $s+1-i$ and $s+1-j$, respectively. This proves (2.3).

The assumption (2.4) implies $a_{ij}^* = a_{ij}$ and $b_i^* = b_i$, so that $\Phi_h^* = \Phi_h$. \square

Explicit Runge–Kutta methods cannot fulfill condition (2.4) with $i = j$, and it is not difficult to see that no explicit Runge–Kutta can be symmetric (Exercise 2). Let us therefore turn our attention to *diagonally implicit Runge–Kutta methods* (DIRK), for which $a_{ij} = 0$ for $i < j$, but with diagonal elements that can be non-zero. In this case condition (2.4) becomes

$$a_{ij} = b_j = b_{s+1-j} \quad \text{for } i > j, \quad a_{jj} + a_{s+1-j, s+1-j} = b_j. \quad (2.6)$$

The Runge–Kutta tableau of such a method is thus of the form (e.g., for $s = 5$)

$$\begin{array}{c|ccccc} c_1 & a_{11} & & & & \\ c_2 & b_1 & a_{22} & & & \\ c_3 & b_1 & b_2 & a_{33} & & \\ 1 - c_2 & b_1 & b_2 & b_3 & a_{44} & \\ 1 - c_1 & b_1 & b_2 & b_3 & b_2 & a_{55} \\ \hline & b_1 & b_2 & b_3 & b_2 & b_1 \end{array} \quad (2.7)$$

with $a_{33} = b_3/2$, $a_{44} = b_2 - a_{22}$, and $a_{55} = b_1 - a_{11}$. If one of the b_i vanishes, then the corresponding stage does not influence the numerical result. This stage can therefore be suppressed, so that the method is equivalent to one with fewer stages. Our next result shows that methods (2.7) can be interpreted as the composition of θ -methods, which are defined as

² For irreducible Runge–Kutta methods, the condition (2.4) is also necessary for symmetry (after a suitable permutation of the stages).

$$\Phi_h^\theta(y_0) = y_1, \quad \text{where} \quad y_1 = y_0 + hf((1-\theta)y_0 + \theta y_1). \quad (2.8)$$

Observe that the adjoint of the θ -method is $\Phi_h^{\theta*} = \Phi_h^{1-\theta}$.

Theorem 2.4. *A diagonally implicit Runge–Kutta method satisfying the symmetry condition (2.4) and $b_i \neq 0$ is equivalent to a composition of θ -methods*

$$\Phi_{b_1 h}^{\alpha_1*} \circ \Phi_{b_2 h}^{\alpha_2*} \circ \dots \circ \Phi_{b_s h}^{\alpha_s} \circ \Phi_{b_1 h}^{\alpha_1}, \quad (2.9)$$

where $\alpha_i = a_{ii}/b_i$.

Proof. Since the θ -method is a Runge–Kutta method with tableau

$$\begin{array}{c|c} \theta & \theta \\ \hline & 1 \end{array}$$

this follows from the discussion in Sect. III.1.3. We have used $\Phi_{b_{s+1-i}h}^{\alpha_{s+1-i}} = \Phi_{b_i h}^{\alpha_i*}$ which holds, because $b_{s+1-i} = b_i$ and $\alpha_{s+1-i} = 1 - \alpha_i$ by (2.6). \square

A more detailed discussion of such methods is therefore postponed to Sect. V.3 on symmetric composition methods.

V.2.2 Partitioned Runge–Kutta Methods

Applying partitioned Runge–Kutta methods (II.2.2) to general partitioned systems

$$\dot{y} = f(y, z), \quad \dot{z} = g(y, z), \quad (2.10)$$

it is obvious that for their symmetry both Runge–Kutta methods have to be symmetric (because $\dot{y} = f(y)$ and $\dot{z} = g(z)$ are special cases of (2.10)). The proof of the following result is identical to that of Theorem 2.3 and therefore omitted.

Theorem 2.5. *If the coefficients of both Runge–Kutta methods b_i, a_{ij} and $\widehat{b}_i, \widehat{a}_{ij}$ satisfy the condition (2.4), then the partitioned Runge–Kutta method (II.2.2) is symmetric.* \square

As a consequence of this theorem we obtain that the Lobatto IIIA–IIIB pair (see Sect. II.2.2) and, in particular, the Störmer–Verlet scheme are symmetric integrators.

An interesting feature of partitioned Runge–Kutta methods is the possibility of having *explicit, symmetric* methods for problems of the form

$$\dot{y} = f(z), \quad \dot{z} = g(y). \quad (2.11)$$

Second order differential equations $\ddot{y} = g(y)$, written in the form $\dot{y} = z, \dot{z} = g(y)$ have this structure, and also all Hamiltonian systems with separable Hamiltonian $H(p, q) = T(p) + V(q)$. It is not possible to get explicit symmetric integrators with non-partitioned Runge–Kutta methods (Exercise 2).

The Störmer–Verlet method (Table II.2.1) applied to (2.11) reads

$$\begin{aligned}
z_{1/2} &= z_0 + \frac{h}{2} g(y_0) \\
y_1 &= y_0 + h f(z_{1/2}) \\
z_1 &= z_{1/2} + \frac{h}{2} g(y_1)
\end{aligned}$$

and is the composition $\Phi_{h/2}^* \circ \Phi_{h/2}$, where

$$\begin{pmatrix} y_1 \\ z_1 \end{pmatrix} = \Phi_h \begin{pmatrix} y_0 \\ z_0 \end{pmatrix}, \quad \begin{aligned} y_1 &= y_0 + h f(z_1) \\ z_1 &= z_0 + h g(y_0) \end{aligned} \quad (2.12)$$

is the symplectic Euler method and

$$\begin{pmatrix} y_1 \\ z_1 \end{pmatrix} = \Phi_h^* \begin{pmatrix} y_0 \\ z_0 \end{pmatrix}, \quad \begin{aligned} y_1 &= y_0 + h f(z_0) \\ z_1 &= z_0 + h g(y_1) \end{aligned} \quad (2.13)$$

its adjoint. All these methods are obviously explicit. How can they be extended to higher order? The idea is to consider partitioned Runge–Kutta methods based on diagonally implicit methods such as in (2.7). If $a_{ii} \cdot \hat{a}_{ii} = 0$, then one component of the i th stage is given explicitly and, due to the special structure of (2.11), the other component is also obtained in a straightforward manner. In order to achieve $a_{ii} \cdot \hat{a}_{ii} = 0$ with a symmetric partitioned method, we have to assume that s , the number of stages, is even.

Theorem 2.6. *A partitioned Runge–Kutta method, based on two diagonally implicit methods satisfying $a_{ii} \cdot \hat{a}_{ii} = 0$ and (2.4) with $b_i \neq 0$ and $\hat{b}_i \neq 0$, is equivalent to a composition of $\Phi_{b_i h}$ and $\Phi_{b_i h}^*$ with Φ_h and Φ_h^* given by (2.12) and (2.13), respectively.* \square

For example, the partitioned method

$$\begin{array}{c|cccc}
& 0 & & & \\
& b_1 & b_2 & & \\
& b_1 & b_2 & 0 & \\
& b_1 & b_2 & b_2 & b_1 \\
\hline
& b_1 & b_2 & b_2 & b_1
\end{array}
\quad
\begin{array}{c|cccc}
& \hat{b}_1 & & & \\
& \hat{b}_1 & 0 & & \\
& \hat{b}_1 & \hat{b}_2 & \hat{b}_2 & \\
& \hat{b}_1 & \hat{b}_2 & \hat{b}_2 & 0 \\
\hline
& \hat{b}_1 & \hat{b}_2 & \hat{b}_2 & \hat{b}_1
\end{array}$$

satisfies the assumptions of the preceding theorem. Since the methods have identical stages, the numerical result only depends on \hat{b}_1 , $b_1 + b_2$, $\hat{b}_2 + \hat{b}_3$, $b_3 + b_4$, and \hat{b}_4 . Therefore, we can assume that $\hat{b}_i = b_i$ and the method is equivalent to the composition $\Phi_{b_1 h}^* \circ \Phi_{b_2 h} \circ \Phi_{b_2 h}^* \circ \Phi_{b_1 h}$.

V.3 Symmetric Composition Methods

In Sect. II.4 the idea of composition methods is introduced, and a systematic way of obtaining high-order methods is outlined. These methods, based on (II.4.4) or on

(II.4.5), turn out to be symmetric, but they require too many stages. A theory of order conditions for general composition methods is developed in Sect. III.3. Here, we apply this theory to the construction of high-order symmetric methods. We mainly follow two lines.

- *Symmetric composition of first order methods.*

$$\Psi_h = \Phi_{\alpha_s h} \circ \Phi_{\beta_s h}^* \circ \dots \circ \Phi_{\beta_2 h}^* \circ \Phi_{\alpha_1 h} \circ \Phi_{\beta_1 h}^*, \quad (3.1)$$

where Φ_h is an arbitrary first order method. In order to make this method symmetric, we assume $\alpha_s = \beta_1$, $\alpha_{s-1} = \beta_2$, etc.

- *Symmetric composition of symmetric methods.*

$$\Psi_h = \Phi_{\gamma_s h} \circ \Phi_{\gamma_{s-1} h} \circ \dots \circ \Phi_{\gamma_2 h} \circ \Phi_{\gamma_1 h}, \quad (3.2)$$

where Φ_h is a symmetric second order method and $\gamma_s = \gamma_1$, $\gamma_{s-1} = \gamma_2$, etc.

V.3.1 Symmetric Composition of First Order Methods

Because of Lemma 3.2 below, every method (3.2) is a special case of method (3.1). In this subsection we concentrate on methods that are of the form (3.1) but not of the form (3.2).

For constructing methods (3.1) of a certain order, one has to solve the system of nonlinear equations given in Theorem III.3.14 (see also Example III.3.15). The symmetry assumption on the coefficients considerably simplifies this system.

Theorem 3.1. *If the coefficients of method (3.1) satisfy $\alpha_{s+1-i} = \beta_i$ for all i , then it is sufficient to consider those trees with odd $\|\tau\|$.*

Proof. This is a consequence of Theorem II.3.2 (the maximal order of symmetric methods is even). In fact, if the condition for order 1 is satisfied, it is automatically of order 2. If, in addition, the conditions for order 3 are satisfied, it is automatically of order 4, etc. \square

It may come as a surprise that the popular leapfrog . . . can be beaten, but only slightly.
(R.I. McLachlan 1995)

Methods of Order 2. The only remaining condition for order two is $\sum_{k=1}^s (\alpha_k + \beta_k) = 1$, and, for $s = 1$, the symmetry requirement leads to $\Phi_{h/2} \circ \Phi_{h/2}^*$. Depending on the choice of Φ_h , this method is equivalent to the midpoint rule, the trapezoidal rule, or the Störmer–Verlet scheme, all very famous and frequently used. However, McLachlan (1995) discovered that the case $s = 2$ can be slightly more advantageous. We obtain

$$\Phi_{\alpha h} \circ \Phi_{(1/2-\alpha)h}^* \circ \Phi_{(1/2-\alpha)h} \circ \Phi_{\alpha h}^*, \quad (3.3)$$

where α is a free parameter, which can serve for clever tuning.

Minimizing the Local Error of Composition Methods. Subtracting the B_∞ -series of the numerical and the exact solutions (see Sect. III.3.2), we obtain

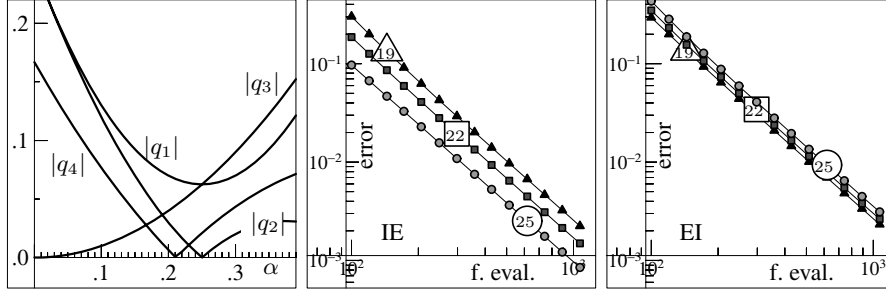


Fig. 3.1. The error functions $|q_i(\alpha)|$ defined in (3.5) (left picture). Work-precision diagrams for the Kepler problem (as in Fig. II.4.4) and for method (3.3) with $\alpha = 0.25$ (Störmer–Verlet), $\alpha = 0.1932$ (McLachlan), and $\alpha = 0.22$. “IE”: method Φ_h treats position by implicit Euler, velocity by explicit Euler; “EI”: method Φ_h treats position by explicit Euler, velocity by implicit Euler

$$\Psi_h(y) - \varphi_h(y) = h^{p+1} \sum_{\|\tau\|=p+1} \frac{1}{\sigma(\tau)} (a(\tau) - e(\tau)) F(\tau)(y) + \mathcal{O}(h^{p+2}).$$

Assuming that the basic method has an expansion $\Phi_h(y) = y + hf(y) + h^2 d_2(y) + h^3 d_3(y) + \dots$, we obtain for method (3.3), similar to (III.3.3), the local error

$$h^3 \left(q_1(\alpha) d_3(y) + q_2(\alpha) (d'_2 f)(y) + q_3(\alpha) (f' d_2)(y) + \frac{1}{2} q_4(\alpha) (f''(f, f))(y) + q_5(\alpha) (f' f' f)(y) \right) + \mathcal{O}(h^4), \quad (3.4)$$

which contains one term for each of the 5 trees $\tau \in T_\infty$ with $\|\tau\| = 3$. The $q_i(\alpha)$ are the polynomials

$$\begin{aligned} q_1(\alpha) &= \frac{1}{4}(1 - 6\alpha + 12\alpha^2), & q_2(\alpha) &= \frac{1}{4}(-1 + 6\alpha - 8\alpha^2), \\ q_3(\alpha) &= -\alpha^2, & q_4(\alpha) &= \frac{1}{6}(1 - 6\alpha + 6\alpha^2), & q_5(\alpha) &= \frac{1}{3}q_1(\alpha), \end{aligned} \quad (3.5)$$

which are plotted in the left picture of Fig. 3.1. If we allow arbitrary basic methods and arbitrary problems, all elementary differentials in the local error are independent, and there is no overall optimal value for α . We see that the modulus of $q_1(\alpha)$ and $q_2(\alpha)$ are minimal for $\alpha = 1/4$, which is precisely the value corresponding to a double application of $\Phi_{h/2} \circ \Phi_{h/2}^*$ with halved step size. But the values $|q_3(\alpha)|$ and $|q_4(\alpha)|$ become smaller with decreasing α (close to $\alpha = 1/4$). McLachlan (1995) therefore minimizes some norm of the error (see Exercise 4) and arrives at the value $\alpha = 0.1932$.

In the numerical experiment of Fig. 3.1 we apply method (3.3) with three different values of α to the Kepler problem (with data as in Fig. II.4.4 and the symplectic Euler method for Φ_h). Once we treat the position variable by the implicit Euler method and the velocity variable by the explicit Euler method (central picture), and

once the other way round (right picture). We notice that the method which is best in one case is worst in the other.

This simple experiment shows that choosing the free parameters of the method by minimizing some arbitrary measure of the error coefficients is problematic. For higher order methods there are many more expressions in the dominating term of the local error (for example: 29 terms for $||\tau|| = 5$). The corresponding functions q_i give a lot of information on the local error, and they indicate the region of parameters that produce good methods. But, unless more information is known about the problem (second order differential equation, nearly integrable systems), one usually minimizes, for orders of 8 or 10, just the maximal values of the α_i , β_i , or γ_i (Kahan & Li 1997).

Methods of Order 4. Theorem 3.1 and Example III.3.15 give 3 conditions for order 4. Therefore, we put $s = 3$ in (3.1) and assume symmetry $\beta_1 = \alpha_3$, $\beta_2 = \alpha_2$, and $\beta_3 = \alpha_1$. This leads to the conditions

$$\alpha_1 + \alpha_2 + \alpha_3 = \frac{1}{2}, \quad \alpha_1^3 + \alpha_2^3 + \alpha_3^3 = 0, \quad (\alpha_3^2 - \alpha_1^2)(\alpha_1 + \alpha_2) = 0.$$

Since with $\alpha_1 + \alpha_2 = 0$ or with $\alpha_1 + \alpha_3 = 0$ the first two of these equations are not compatible, the unique solution of this system is

$$\alpha_1 = \alpha_3 = \frac{1}{2(2 - 2^{1/3})}, \quad \alpha_2 = -\frac{2^{1/3}}{2(2 - 2^{1/3})}.$$

We observe that $\beta_i = \alpha_i$ for all i . Therefore, $\Phi_{\alpha_i h} \circ \Phi_{\beta_i h}^*$ can be grouped together in (3.1) and we have obtained a method of type (3.2), which is actually method (II.4.4) with $p = 2$.

Again, the solutions with the minimal number of stages do not give the best methods (remember the good performance of Suzuki's fourth order method (II.4.5) in Fig. II.4.4), so we look for 4th order methods with larger s . McLachlan (1995) has constructed a method for $s = 5$ with particularly small error terms and nice coefficients

$$\begin{aligned} \beta_1 = \alpha_5 &= \frac{14 - \sqrt{19}}{108}, & \alpha_1 = \beta_5 &= \frac{146 + 5\sqrt{19}}{540}, \\ \beta_2 = \alpha_4 &= \frac{-23 - 20\sqrt{19}}{270}, & \alpha_2 = \beta_4 &= \frac{-2 + 10\sqrt{19}}{135}, & \beta_3 = \alpha_3 &= \frac{1}{5}, \end{aligned} \quad (3.6)$$

which he recommends “for all uses”.

In Fig. 3.2 we compare the numerical performances of all these methods on our already well-known example in both variants (implicit-explicit and vice-versa). We see that the best methods in *one* picture may be worse in the other. For comparison, the results are surrounded by “ghosts in grey” representing good formulae from the next lower (order 2) and the next higher (order 6) class of methods.

Methods Tuned for Special Problems. In the case where one is applying a *special* method to a *special* problem (e.g., to second order differential equations or to small

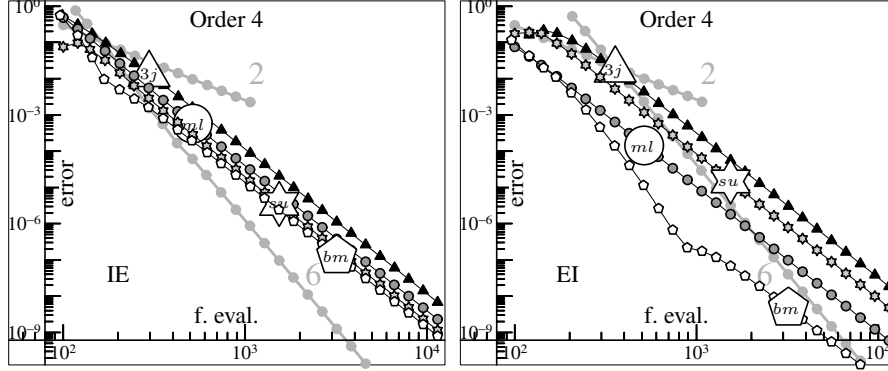


Fig. 3.2. Work-precision diagrams for methods of order 4 as in Fig. 3.1; “3j”: the Triple Jump (II.4.4); “su”: method (II.4.5) of Suzuki; “ml”: McLachlan (3.6); “bm”: method (3.7); in grey: neighbouring order methods Störmer/Verlet (order 2) and $p_6 s_9$ (order 6)

perturbations of integrable systems), more spectacular gains of efficiency are possible. For example, Blanes & Moan (2002) have constructed the following fourth order method with $s = 6$

$$\begin{aligned} \beta_1 = \alpha_6 &= 0.082984406417405, & \alpha_1 = \beta_6 &= 0.16231455076687, \\ \beta_2 = \alpha_5 &= 0.23399525073150, & \alpha_2 = \beta_5 &= 0.37087741497958, \\ \beta_3 = \alpha_4 &= -0.40993371990193, & \alpha_3 = \beta_4 &= 0.059762097006575, \end{aligned} \quad (3.7)$$

which, when correctly applied to second order differential equations (right picture of Fig. 3.2) exhibits excellent performance.

Further methods, adapted to the integration of second order differential equations, have been constructed by Forest (1992), McLachlan & Atela (1992), Calvo & Sanz-Serna (1993), Okunbor & Skeel (1994), and McLachlan (1995). Another important situation, which allows a tuning of the parameters, are near-integrable systems such as the perturbed two-body motion (e.g., the outer solar system considered in Chap. I). If the differential equation can be split into $\dot{y} = f^{[1]}(y) + f^{[2]}(y)$, where $\dot{y} = f^{[1]}(y)$ is exactly integrable and $f^{[2]}(y)$ is small compared to $f^{[1]}(y)$, special integrators should be used. We refer to Kinoshita, Yoshida & Nakai (1991), Wisdom & Holman (1991), Saha & Tremaine (1992), and McLachlan (1995b) for more details and for the parameters of such integrators.

Methods of Order 6. By Theorem 3.1 and Example III.3.12 a method (3.1) has to satisfy 9 conditions for order 6. It turns out that these order conditions have already a solution with $s = 7$, for all known solutions with $s \leq 8$ are equivalent to methods of type (3.2). With order 6 we are apparently close to the point where the enormous simplifications of the order conditions due to Theorem 3.3 below start to outperform the freedom of choosing different values for α_i and β_i . We therefore continue our discussion by considering only the special case (3.2).

V.3.2 Symmetric Composition of Symmetric Methods

The introduction of more symmetries into the method simplifies considerably the order conditions. These simplifications can be best understood with a sort of “Choleski decomposition” of symmetric methods (Murua & Sanz-Serna 1999).

Lemma 3.2. *For every symmetric method $\Phi_h(y)$ that admits an expansion in powers of h , there exists $\widehat{\Phi}_h(y)$ such that*

$$\Phi_h(y) = (\widehat{\Phi}_{h/2} \circ \widehat{\Phi}_{h/2}^*)(y).$$

Proof. Since $\Phi_h(y) = y + \mathcal{O}(h)$ is close to the identity, the existence of a unique method $\widehat{\Phi}_h(y) = y + hd_1(y) + h^2d_2(y) + \dots$ satisfying $\Phi_h = \widehat{\Phi}_{h/2} \circ \widehat{\Phi}_{h/2}$ follows from Taylor expansion and from a comparison of like powers of h .

If $\Phi_h(y)$ is symmetric, we have in addition

$$\Phi_h = \Phi_h^{-1} = \widehat{\Phi}_{-h/2}^{-1} \circ \widehat{\Phi}_{-h/2}^{-1},$$

and $\widehat{\Phi}_{h/2} = \widehat{\Phi}_{-h/2}^{-1} = \widehat{\Phi}_{h/2}^*$ follows from the uniqueness of $\widehat{\Phi}_h$. \square

We let Φ_h be a symmetric method, and we consider the composition

$$\Psi_h = \Phi_{\gamma_s h} \circ \dots \circ \Phi_{\gamma_2 h} \circ \Phi_{\gamma_1 h}. \quad (3.8)$$

Using the method $\widehat{\Phi}_h$ of Lemma 3.2, this composition method is equivalent to (3.1) (Φ_h replaced with $\widehat{\Phi}_h$) with

$$\alpha_i = \beta_i = \frac{\gamma_i}{2}. \quad (3.9)$$

Theorem 3.3. *For composition methods (3.8) with symmetric Φ_h it is sufficient to consider the order conditions of Theorem III.3.14 for $\tau \in \mathcal{H}$ where all vertices of τ have odd indices.*

Proof. If $i(\tau)$ is even, it follows from $\alpha_k = \beta_k$ and from (III.3.11) that

$$a_s(\tau) = a_{s-1}(\tau) = \dots = a_1(\tau) = a_0(\tau) = 0.$$

Since $e(\tau) = 0$ for such trees, the corresponding order condition is automatically satisfied. Any other vertex with an even index can be brought to the root by applying the Switching Lemma III.3.8. \square

After this reduction, only 7 conditions survive for order 6 from the trees displayed in Example III.3.12. A further reduction in the number of order conditions is achieved by assuming *symmetric coefficients* in method (3.8), i.e.,

$$\gamma_{s+1-j} = \gamma_j \quad \text{for all } j. \quad (3.10)$$

This implies that the overall method Ψ_h is symmetric, so that the order conditions for trees with an even $\|\tau\|$ need not be considered. This proves the following result.

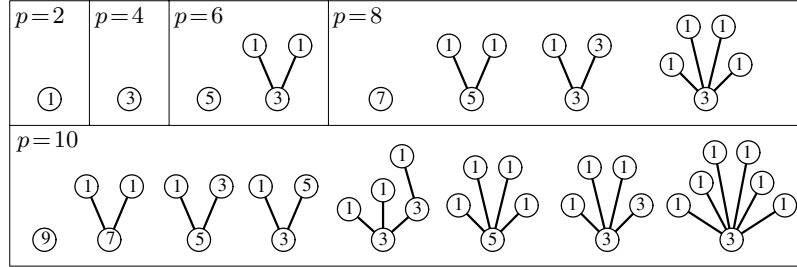






Fig. 3.3. Symmetric Composition of Symmetric Methods up to order 10

Theorem 3.4. For composition methods (3.8) with symmetric Φ_h , satisfying (3.10), it is sufficient to consider the order conditions for $\tau \in \mathcal{H}$ where all vertices of τ have odd indices and where $\|\tau\|$ is odd. \square

Figure 3.3 shows the remaining order conditions for methods up to order 10. We see that for order 6 there remain only 4 conditions, much less than the 166 that we started with (Theorem III.3.6).

Example 3.5. The rule of (III.3.14) leads to the following conditions for *symmetric* composition of *symmetric* methods:

Order 2:	①	$\sum_{k=1}^s \gamma_k = 1$	
Order 4:	③	$\sum_{k=1}^s \gamma_k^3 = 0$	
Order 6:	⑤	$\sum_{k=1}^s \gamma_k^5 = 0$	 $\sum_{k=1}^s \gamma_k^3 \left(\sum_{\ell=1}^k \gamma_\ell \right)^2 = 0$
Order 8:	⑦	$\sum_{k=1}^s \gamma_k^7 = 0$	 $\sum_{k=1}^s \gamma_k^5 \left(\sum_{\ell=1}^k \gamma_\ell \right)^2 = 0$
		 $\sum_{k=1}^s \gamma_k^3 \sum_{\ell=1}^k \gamma_\ell \sum_{m=1}^k \gamma_m^3 = 0$	 $\sum_{k=1}^s \gamma_k^3 \left(\sum_{\ell=1}^k \gamma_\ell \right)^4 = 0.$

Here, similar to Example III.3.15, a *prime* attached to a summation symbol indicates that the last term γ_ℓ^i is taken as $\gamma_\ell^i/2$.

Methods of Order 4. The methods (II.4.4) and (II.4.5) are both of the form (3.8), and those with $p = 2$ yield methods of order 4. We have seen in the experiment of Fig. II.4.4 that the method (II.4.5) yields more precise approximations; see also Fig. 3.2. We do not know of any 4th order method of type (3.2) that is significantly better than method (3.1) with coefficients (3.6).

Methods of Order 6. If we search for a minimal stage solution of the four equations for order 6, we apparently need four free parameters $\gamma_1, \gamma_2, \gamma_3, \gamma_4$; then $\gamma_5, \gamma_6, \gamma_7$ are determined by symmetry. Equation ① gives $\gamma_4 = 1 - 2(\gamma_1 + \gamma_2 + \gamma_3)$. So we end up with three equations for the three unknowns $\gamma_1, \gamma_2, \gamma_3$. A numerical search for this problem produces three solutions, the best of which has been discovered by many authors, in particular by Yoshida (1990), and is as follows:

$$\begin{aligned} \gamma_1 = \gamma_7 &= 0.78451361047755726381949763 \\ \gamma_2 = \gamma_6 &= 0.23557321335935813368479318 \\ \gamma_3 = \gamma_5 &= -1.17767998417887100694641568 \\ \gamma_4 &= 1.31518632068391121888424973 \end{aligned} \quad \begin{array}{c} p6 \ s7 \\ \text{Diagram showing 7 stages} \end{array} \quad (3.11)$$

Using computer algebra, Koseleff (1996) proves that the nonlinear system for $\gamma_1, \gamma_2, \gamma_3$ has not more than three real solutions.

Similar to the situation for order 4, where relaxing the minimal number of stages allowed a significant increase of performance, we also might expect to obtain better methods of order 6 in this way. McLachlan (1995) increases s by two and constructs good methods with small error coefficients. By minimizing $\max_i |\gamma_i|$, Kahan & Li (1997) obtain the following excellent method³

$$\begin{aligned} \gamma_1 = \gamma_9 &= 0.39216144400731413927925056 \\ \gamma_2 = \gamma_8 &= 0.33259913678935943859974864 \\ \gamma_3 = \gamma_7 &= -0.70624617255763935980996482 \\ \gamma_4 = \gamma_6 &= 0.08221359629355080023149045 \\ \gamma_5 &= 0.79854399093482996339895035 \end{aligned} \quad \begin{array}{c} p6 \ s9 \\ \text{Diagram showing 9 stages} \end{array} \quad (3.12)$$

This method produces, with a comparable number of total steps, errors which are typically smaller than those of method (3.11). Numerical results of these two methods are given in Fig. 3.4.

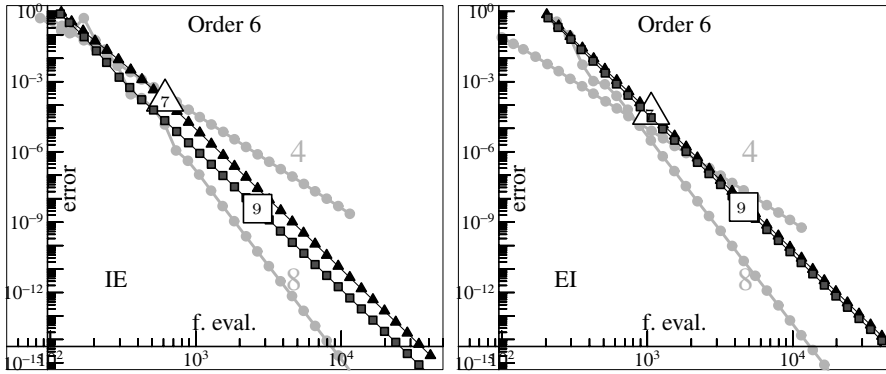
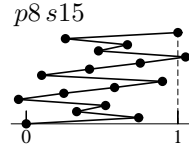


Fig. 3.4. Work-precision diagrams for methods of order 6 for the Kepler problem as in Fig. 3.1; “7”: method $p6 \ s7$ of (3.11); “9”: method $p6 \ s9$ of (3.12); in grey: neighbouring order methods (3.6) (order 4) and $p8 \ s17$ (order 8)

³ The authors are grateful to S. Blanes for this reference.

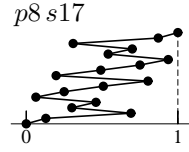
Methods of Order 8. For order 8, Fig. 3.3 represents 8 equations to solve. This indicates that the minimal value of s is 15. A numerical search for solutions $\gamma_1, \dots, \gamma_8$ of these equations produces hundreds of solutions. We choose among all these the solution with the smallest $\max(|\gamma_i|)$. The coefficients, which were originally given by Suzuki & Umeno (1993), Suzuki (1994), and later by McLachlan (1995), are as follows:

$$\begin{aligned}
 \gamma_1 = \gamma_{15} &= 0.74167036435061295344822780 \\
 \gamma_2 = \gamma_{14} &= -0.40910082580003159399730010 \\
 \gamma_3 = \gamma_{13} &= 0.19075471029623837995387626 \\
 \gamma_4 = \gamma_{12} &= -0.57386247111608226665638773 \\
 \gamma_5 = \gamma_{11} &= 0.29906418130365592384446354 \\
 \gamma_6 = \gamma_{10} &= 0.33462491824529818378495798 \\
 \gamma_7 = \gamma_9 &= 0.31529309239676659663205666 \\
 \gamma_8 &= -0.79688793935291635401978884
 \end{aligned}
 \tag{3.13}$$



By putting $s = 17$ we obtain one degree of freedom in solving the equations. This allows an improvement on the foregoing method. The best known solution, slightly better than a method of McLachlan (1995), has been found by Kahan & Li (1997) and is given by

$$\begin{aligned}
 \gamma_1 = \gamma_{17} &= 0.13020248308889008087881763 \\
 \gamma_2 = \gamma_{16} &= 0.56116298177510838456196441 \\
 \gamma_3 = \gamma_{15} &= -0.38947496264484728640807860 \\
 \gamma_4 = \gamma_{14} &= 0.15884190655515560089621075 \\
 \gamma_5 = \gamma_{13} &= -0.39590389413323757733623154 \\
 \gamma_6 = \gamma_{12} &= 0.18453964097831570709183254 \\
 \gamma_7 = \gamma_{11} &= 0.25837438768632204729397911 \\
 \gamma_8 = \gamma_{10} &= 0.29501172360931029887096624 \\
 \gamma_9 &= -0.60550853383003451169892108
 \end{aligned}
 \tag{3.14}$$



Numerical results, in the same style as above, are given in Fig. 3.5.

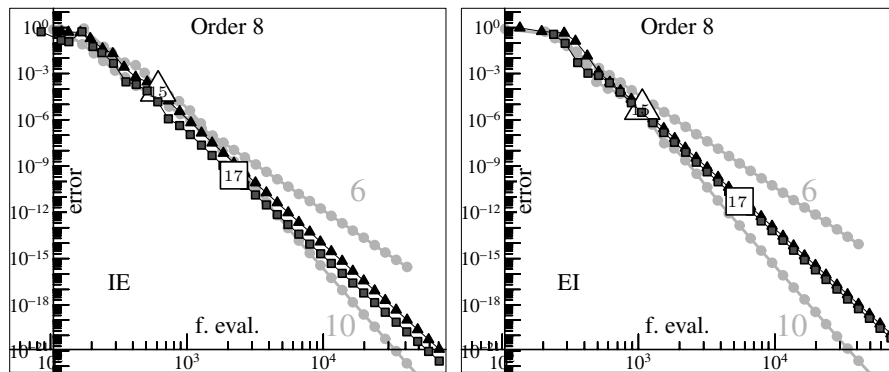
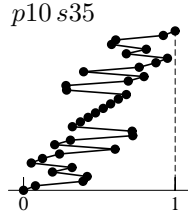


Fig. 3.5. Work-precision diagrams for methods of order 8 for the Kepler problem as in Fig. 3.1; “15”: method $p8\ s15$ of (3.13); “17”: method $p8\ s17$ of (3.14); in grey: neighbouring order methods $p6\ s9$ (order 6) and $p10\ s35$ (order 10)

Methods of Order 10. The first methods of order 10 were given by Kahan & Li (1997) with $s = 31$ and $s = 33$, which could be improved on after some nights of computer search (see method (V.3.15) of the first edition). A significantly improved method for $s = 35$ (see Fig. 3.5 for a comparison with eighth order methods) has in the meantime been found by Sofroniou & Spaletta (2004):

$$\begin{aligned}
 \gamma_1 = \gamma_{35} &= 0.07879572252168641926390768 \\
 \gamma_2 = \gamma_{34} &= 0.31309610341510852776481247 \\
 \gamma_3 = \gamma_{33} &= 0.02791838323507806610952027 \\
 \gamma_4 = \gamma_{32} &= -0.22959284159390709415121340 \\
 \gamma_5 = \gamma_{31} &= 0.13096206107716486317465686 \\
 \gamma_6 = \gamma_{30} &= -0.26973340565451071434460973 \\
 \gamma_7 = \gamma_{29} &= 0.07497334315589143566613711 \\
 \gamma_8 = \gamma_{28} &= 0.11199342399981020488957508 \\
 \gamma_9 = \gamma_{27} &= 0.36613344954622675119314812 \\
 \gamma_{10} = \gamma_{26} &= -0.39910563013603589787862981 \\
 \gamma_{11} = \gamma_{25} &= 0.10308739852747107731580277 \\
 \gamma_{12} = \gamma_{24} &= 0.41143087395589023782070412 \\
 \gamma_{13} = \gamma_{23} &= -0.00486636058313526176219566 \\
 \gamma_{14} = \gamma_{22} &= -0.39203335370863990644808194 \\
 \gamma_{15} = \gamma_{21} &= 0.05194250296244964703718290 \\
 \gamma_{16} = \gamma_{20} &= 0.05066509075992449633587434 \\
 \gamma_{17} = \gamma_{19} &= 0.04967437063972987905456880 \\
 \gamma_{18} &= 0.04931773575959453791768001
 \end{aligned}
 \tag{3.15}$$



V.3.3 Effective Order and Processing Methods

There has recently been a revival of interest in the concept of “effective order”.
(J.C. Butcher 1998)

The concept of effective order was introduced by Butcher (1969) with the aim of constructing 5th order explicit Runge–Kutta methods with 5 stages. The idea is to search for a computationally efficient method K_h such that with a suitable χ_h ,

$$\Psi_h = \chi_h \circ K_h \circ \chi_h^{-1} \tag{3.16}$$

has an order higher than that of K_h . The method K_h is called the *kernel*, and χ_h can be interpreted as a transformation in the phase space, close to the identity. Because of

$$\Psi_h^N = \chi_h \circ K_h^N \circ \chi_h^{-1},$$

an implementation of Ψ_h over N steps with constant step size h has the same computational efficiency as K_h . The computation of χ_h^{-1} has only to be done once at the beginning of the integration, and χ_h has to be evaluated only at output points, which can be performed on another processor. In the article López-Marcos, Sanz-Serna & Skeel (1996) the notion of *preprocessing* for the step χ_h^{-1} and *postprocessing* for χ_h is introduced.

Example 3.6 (Störmer–Verlet as Processed Symplectic Euler Method). Consider a split differential equation, let $\Phi_h^{[LT]} = \varphi_h^{[2]} \circ \varphi_h^{[1]}$ be the Lie–Trotter formula or symplectic Euler method (see Sect. II.5), and $\Phi_h^{[S]} = \varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}$ the Strang splitting or Störmer–Verlet scheme. As a consequence of the group property of the exact flow, we have

$$\Phi_h^{[S]} = \varphi_{h/2}^{[1]} \circ \Phi_h^{[LT]} \circ \varphi_{h/2}^{[1]} = \chi_h \circ \Phi_h^{[LT]} \circ \chi_h^{-1}$$

with $\chi_h = \varphi_{h/2}^{[1]}$. Hence, applying the Lie–Trotter formula with processing yields a second order approximation.

Since the use of geometric integrators requires constant step sizes, it is quite natural that Butcher’s idea of effective order has been revived in this context. A systematic search for processed composition methods started with the works of Wisdom, Holman & Touma (1996), McLachlan (1996), and Blanes, Casas & Ros (1999, 2000b).

Let us explain the technique of processing in the situation where the kernel K_h is a symmetric composition

$$K_h = \Phi_{\gamma_s h} \circ \dots \circ \Phi_{\gamma_2 h} \circ \Phi_{\gamma_1 h} \quad (\gamma_{s+1-i} = \gamma_i \text{ for all } i) \quad (3.17)$$

of a symmetric method Φ_h . We suppose that the processor is of the form

$$\chi_h = \Phi_{\delta_r h} \circ \dots \circ \Phi_{\delta_2 h} \circ \Phi_{\delta_1 h}, \quad (3.18)$$

such that its inverse is given by (use the symmetry $\Phi_h^{-1} = \Phi_{-h}$)

$$\chi_h^{-1} = \Phi_{-\delta_1 h} \circ \Phi_{-\delta_2 h} \circ \dots \circ \Phi_{-\delta_r h}. \quad (3.19)$$

Order Conditions. The composite method $\Psi_h = \chi_h \circ K_h \circ \chi_h^{-1}$ is of the form $\Psi_h = \Phi_{\varepsilon_{2r+s} h} \circ \dots \circ \Phi_{\varepsilon_2 h} \circ \Phi_{\varepsilon_1 h}$ with

$$(\varepsilon_{2r+s}, \dots, \varepsilon_2, \varepsilon_1) = (\delta_r, \dots, \delta_1, \gamma_s, \dots, \gamma_1, -\delta_1, \dots, -\delta_r). \quad (3.20)$$

Theorem 3.3 thus tells us that only the order conditions corresponding to $\tau \in \mathcal{H}$, whose vertices have odd indices, have to be considered. Unfortunately, the sequence $\{\varepsilon_i\}$ of (3.20) does not satisfy the symmetry relation (3.10), unless all δ_i vanish. However, if we require

$$\chi_{-h}(y) = \chi_h(y) + \mathcal{O}(h^{p+1}), \quad (3.21)$$

we see that $\chi_h^{-1}(y) = \chi_h^*(y) + \mathcal{O}(h^{p+1})$, and the method $\Psi_h = \chi_h \circ K_h \circ \chi_h^{-1}$ is symmetric up to terms of order $\mathcal{O}(h^{p+1})$. Consequently, the reduction of Theorem 3.4 is valid, so that for order p only the trees of Fig. 3.3 have to be considered.

For the first tree of Example 3.5 the order condition is

$$1 = \sum_{k=1}^{2r+s} \varepsilon_k = \sum_{k=1}^s \gamma_k,$$

and we see that this is a condition on the kernel K_h only. Similarly, for odd i we have

$$0 = \sum_{k=1}^{2r+s} \varepsilon_k^i = \sum_{k=1}^s \gamma_k^i, \quad (3.22)$$

so that also the trees ③, ⑤, ⑦, ... give conditions on K_h and cannot be influenced by the processor. We next consider the trees of Example 3.5 with three vertices, whose order condition is

$$0 = \sum_{k=1}^{2r+s} \varepsilon_k^i \sum_{\ell=1}^k \varepsilon_\ell^j \sum_{m=1}^k \varepsilon_m^q.$$

We split the sums according to the partitioning into $\delta_i, \gamma_i, -\delta_i$ in (3.20), and we denote the expressions appearing in Example 3.5 by $a(\tau)$ and those corresponding to χ_h and χ_h^{-1} by $b(\tau)$ and $b^{-1}(\tau)$, respectively. Using the abbreviations τ_i for the tree with one vertex labelled i , τ_{ij} for the tree with two vertices labelled i (the root) and j , and by τ_{ijq} the trees with three vertices labelled i (root), j and q (vertices that are directly connected to the root), this yields

$$\begin{aligned} 0 = & b^{-1}(\tau_{ijq}) + a(\tau_i)b^{-1}(\tau_j)b^{-1}(\tau_q) + a(\tau_{ij})b^{-1}(\tau_q) \\ & + a(\tau_{iq})b^{-1}(\tau_j) + a(\tau_{ijq}) + b(\tau_i)b^{-1}(\tau_j)b^{-1}(\tau_q) \\ & + b(\tau_i)b^{-1}(\tau_j)a(\tau_q) + b(\tau_i)a(\tau_j)b^{-1}(\tau_q) + b(\tau_i)a(\tau_j)a(\tau_q) \\ & + b(\tau_{ij})b^{-1}(\tau_q) + b(\tau_{ij})a(\tau_q) + b(\tau_{iq})b^{-1}(\tau_j) + b(\tau_{iq})a(\tau_j) + b(\tau_{ijq}). \end{aligned} \quad (3.23)$$

How can we simplify this long expression? First of all, we imagine K_h to be the identity (either $s = 0$ or all $\gamma_i = 0$), so that $\Psi_h = \chi_h \circ \chi_h^{-1}$ becomes the identity. In this situation, the terms involving $a(\tau)$ are not present in (3.23), and we obtain

$$0 = b^{-1}(\tau_{ijq}) + b(\tau_i)b^{-1}(\tau_j)b^{-1}(\tau_q) + b(\tau_{ij})b^{-1}(\tau_q) + b(\tau_{iq})b^{-1}(\tau_j) + b(\tau_{ijq}).$$

We can thus remove all terms in (3.23) that do not contain a factor $a(\tau)$. Now observe that by (3.21), $\chi_h(y)$ as well as $\chi_h^{-1}(y)$ have an expansion in even powers of h . Therefore, $b(\tau)$ and $b^{-1}(\tau)$ vanish for all τ with odd $\|\tau\|$. Formula (3.23) thus simplifies considerably and yields

$$0 = a(\tau_{311}) + 2b(\tau_{31})a(\tau_1), \quad (3.24)$$

$$0 = a(\tau_{511}) + 2b(\tau_{51})a(\tau_1), \quad (3.25)$$

$$0 = a(\tau_{313}) + b(\tau_{31})a(\tau_3) + b(\tau_{33})a(\tau_1). \quad (3.26)$$

A similar computation for the last tree in Example 3.5 gives (in an obvious notation)

$$0 = a(\tau_{31111}) + 4b(\tau_{31})a(\tau_1)^3 + 4b(\tau_{311})a(\tau_1). \quad (3.27)$$

Since $a(\tau_1) = \sum_{i=1}^s \gamma_i = 1$, the conditions (3.24), (3.25) and (3.27) can be interpreted as conditions on the processor, namely on $b(\tau_{31})$, $b(\tau_{51})$ and $b(\tau_{3111})$. We

already have $a(\tau_3) = 0$ from (3.22), and an application of the Switching Lemma III.3.8 gives $b(\tau_{33}) = \frac{1}{2}(b(\tau_3)^2 - b(\tau_6))$. The term $b(\tau_3)$ vanishes by (3.21) and $b(\tau_6) = 0$ is a consequence of the proof of Theorem 3.3. Therefore (3.26) is equivalent to $a(\tau_{313}) = 0$. We summarize our computation in the following theorem.

Theorem 3.7. *The processing method $\Psi_h = \chi_h \circ K_h \circ \chi_h^{-1}$ is of order p ($p \leq 8$), if*

- *the coefficients γ_i of the kernel satisfy the conditions of the left column in Example 3.5, i.e., 3 conditions for order 6, and 5 conditions for order 8;*
- *the coefficients δ_i of the processor are such that (3.21) holds (4 conditions for order 6, and 8 conditions for order 8), and in addition condition (3.24) for order 6, and (3.24), (3.25), (3.27) for order 8 are satisfied.* \square

Remark 3.8. Although we have presented the computations only for $p \leq 8$, the result is general. All trees $\tau \in \mathcal{H}$, which are not of the form $\tau = u \circ \textcircled{1}$, give rise to conditions on the kernel K_h (for a similar result in the context of Runge–Kutta methods see Butcher & Sanz-Serna (1996)). The remaining conditions have to be satisfied by the coefficients of the processor. Due to the reduced number of order conditions, it is relatively easy to construct high order kernels. However, the difficulty in constructing a suitable processor increases rapidly with the order.

The application of the processing technique is two-fold. A first possibility is to take one of the high-order composition methods of the form (3.2), e.g., one of those presented in Sect. V.3.2, and to exploit the freedom in the coefficients of the processor to make the error constants smaller.

Another possibility is to start from the beginning and to construct a method K_h with coefficients satisfying only the conditions of Theorem 3.7. Methods of effective order 6 and 8 have been constructed in this way by Blanes (2001).

V.4 Symmetric Methods on Manifolds

Numerical methods for differential equations on manifolds have been introduced in Sections IV.4 and IV.5. The presented algorithms are in general not symmetric. We discuss here suitable symmetric modifications which often have an improved long-time behaviour. We consider a differential equation

$$\dot{y} = f(y), \quad f(y) \in T_y \mathcal{M} \quad (4.1)$$

on a manifold \mathcal{M} , and we assume that the manifold is either given as the zero set of a function $g(y)$ or by means of a suitable parametrization $y = \varphi(z)$.

V.4.1 Symmetric Projection

Due to the projection at the end of an integration step, the standard projection method (Algorithm IV.4.2) is not symmetric (see Fig. IV.4.2). In order to make the

overall algorithm symmetric, one has to apply a kind of “inverse projection” at the beginning of each integration step. This idea has first been used by Ascher & Reich (1999) to enforce conservation of energy, and it has been applied in more general contexts by Hairer (2000).

Algorithm 4.1 (Symmetric Projection Method). Assume that $y_n \in \mathcal{M}$. One step $y_n \mapsto y_{n+1}$ is defined as follows (see Fig. 4.1, right picture):

- $\tilde{y}_n = y_n + G(y_n)^T \mu$ where $g(y_n) = 0$ (perturbation step);
- $\tilde{y}_{n+1} = \Phi_h(\tilde{y}_n)$ (symmetric one-step method applied to $\dot{y} = f(y)$);
- $y_{n+1} = \tilde{y}_{n+1} + G(y_{n+1})^T \mu$ with μ such that $g(y_{n+1}) = 0$ (projection step).

Here, $G(y) = g'(y)$ denotes the Jacobian of $g(y)$. It is important to take a symmetric method in the second step, and the same vector μ in the perturbation and projection steps.

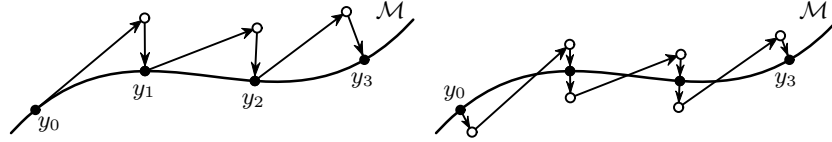


Fig. 4.1. Standard projection (left picture) compared to symmetric projection (right)

Existence of the Numerical Solution. The vector μ and the numerical approximation y_{n+1} are implicitly defined by

$$F(h, y_{n+1}, \mu) = \begin{pmatrix} y_{n+1} - \Phi_h(y_n + G(y_n)^T \mu) - G(y_{n+1})^T \mu \\ g(y_{n+1}) \end{pmatrix} = 0. \quad (4.2)$$

Since $F(0, y_n, 0) = 0$ and since

$$\frac{\partial F}{\partial(y_{n+1}, \mu)}(0, y_n, 0) = \begin{pmatrix} I & -2G(y_n)^T \\ G(y_n) & 0 \end{pmatrix} \quad (4.3)$$

is invertible (provided that $G(y_n)$ has full rank), an application of the implicit function theorem proves the existence of the numerical solution for sufficiently small step size h . The simple structure of the matrix (4.3) can also be exploited for an efficient solution of the nonlinear system (4.2) using simplified Newton iterations. If the basic method Φ_h is itself implicit, the nonlinear system (4.2) should be solved in tandem with $\tilde{y}_{n+1} = \Phi_h(\tilde{y}_n)$.

Order. For a study of the local error we let $y_n := y(t_n)$ be a value on the exact solution $y(t)$ of (4.1). If the basic method Φ_h is of order p , i.e., if $y(t_n + h) - \Phi_h(y(t_n)) = \mathcal{O}(h^{p+1})$, we have $F(h, y(t_{n+1}), 0) = \mathcal{O}(h^{p+1})$. Compared to (4.2) the implicit function theorem yields

$$y_{n+1} - y(t_{n+1}) = \mathcal{O}(h^{p+1}) \quad \text{and} \quad \mu = \mathcal{O}(h^{p+1}).$$

This proves that the symmetric projection method of Algorithm 4.1 has the same order as the underlying one-step method Φ_h .

Symmetry of the Algorithm. Exchanging $h \leftrightarrow -h$ and $y_n \leftrightarrow y_{n+1}$ in the Algorithm 4.1 yields

$$\begin{aligned} \tilde{y}_n &= y_{n+1} + G(y_{n+1})^T \mu, & g(y_{n+1}) &= 0, \\ \tilde{y}_{n+1} &= \Phi_{-h}(\tilde{y}_n), \\ y_n &= \tilde{y}_{n+1} + G(y_n)^T \mu, & g(y_n) &= 0. \end{aligned}$$

The auxiliary variables μ , \tilde{y}_n , and \tilde{y}_{n+1} can be arbitrarily renamed. If we replace them with $-\mu$, \tilde{y}_{n+1} , and \tilde{y}_n , respectively, we get the formulas of the original algorithm provided that the method Φ_h of the intermediate step is symmetric. This proves the symmetry of the algorithm.

Various modifications of the perturbation and projection steps are possible without destroying the symmetry. For example, one can replace the arguments y_n and y_{n+1} in $G(y)$ with $(y_n + y_{n+1})/2$. It might be advantageous to use a constant direction, i.e., $\tilde{y}_n = y_n + A^T \mu$, $y_{n+1} = \tilde{y}_{n+1} + A^T \mu$ with a constant matrix A . In this case the matrix $G(y)A^T$ has to be invertible along the solution in order to guarantee the existence of the numerical solution.

Reversibility. From Theorem 1.5 we know that symmetry alone does not imply the ρ -reversibility of the numerical flow. The method must also satisfy the compatibility condition (1.4). It is straightforward to check that this condition is satisfied if the integrator Φ_h of the intermediate step of Algorithm 4.1 satisfies (1.4) and, in addition,

$$\rho G(y)^T = G(\rho y)^T \sigma \quad (4.4)$$

holds with some constant invertible matrix σ . In many interesting situations we have $g(\rho y) = \sigma^{-T} g(y)$ with a suitable σ , so that (4.4) follows by differentiation if $\rho \rho^T = I$. Similarly, when a projection with constant direction $y = \tilde{y} + A^T \mu$ is applied, the matrix A has to satisfy $\rho A^T = A^T \sigma$ for a suitably chosen invertible matrix σ (see the experiment of Example 4.4 below).

Example 4.2. Let us consider the equations of motion of a rigid body as described in Example IV.1.7. They constitute a differential equation on the manifold

$$\mathcal{M} = \{(y_1, y_2, y_3) \mid y_1^2 + y_2^2 + y_3^2 = 1\},$$

and it is ρ -reversible with respect to $\rho(y_1, y_2, y_3) = (-y_1, y_2, y_3)$, and also with respect to $\rho(y_1, y_2, y_3) = (y_1, -y_2, y_3)$ and $\rho(y_1, y_2, y_3) = (y_1, y_2, -y_3)$. For a numerical simulation we take $I_1 = 2$, $I_2 = 1$, $I_3 = 2/3$, and the initial value $y_0 = (\cos(0.9), 0, \sin(0.9))$. We apply the trapezoidal rule (II.1.2) with the large step size $h = 1$ in three different versions.

The upper picture of Fig. 4.2 shows the result of a direct application of the trapezoidal rule. The numerical solution lies apparently on a closed curve, but it does not

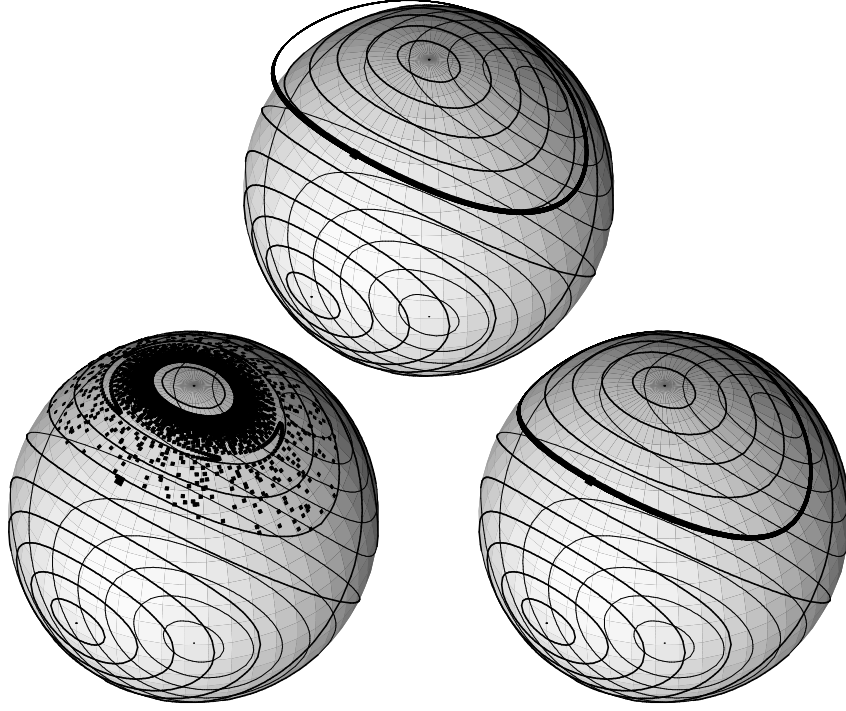


Fig. 4.2. Numerical simulation of the rigid body equations. The three pictures correspond to a direct application (upper), to the standard projection (lower left), and to the symmetric projection (lower right) of the trapezoidal rule; 5000 steps with step size $h = 1$

lie exactly on the manifold \mathcal{M} . This can be seen as follows: the trapezoidal rule Φ_h^T is conjugate to the implicit midpoint rule Φ_h^M via a half-step of the explicit Euler method $\chi_{h/2}$. In fact the relations

$$\Phi_h^T = \chi_{h/2}^* \circ \chi_{h/2} \quad \text{and} \quad \Phi_h^M = \chi_{h/2} \circ \chi_{h/2}^*$$

hold, so that

$$\Phi_h^T = \chi_{h/2}^{-1} \circ \Phi_h^M \circ \chi_{h/2} \quad \text{and} \quad (\Phi_h^T)^N = \chi_{h/2}^{-1} \circ (\Phi_h^M)^N \circ \chi_{h/2}.$$

Consequently, the trajectory of the trapezoidal rule is obtained from the trajectory of the midpoint rule by a simple change of coordinates. On the other hand, the numerical solution of the midpoint rule lies exactly on a solution curve because it conserves quadratic invariants (Theorem IV.2.1).

Using standard orthogonal projection (Algorithm IV.4.2) we obviously obtain a numerical solution lying on the manifold \mathcal{M} . But as we can see from the lower left picture of Fig. 4.2, it does not remain near a closed curve and converges to a fixed point. The lower right picture shows that the use of the symmetric orthogonal projection (Algorithm 4.1) recovers the property of remaining near the closed solution curve.

Example 4.3 (Numerical Experiment with Constant Direction of Projection).

We consider the pendulum equation in Cartesian coordinates (see Example IV.4.1),

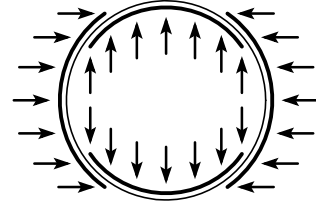
$$\begin{aligned}\dot{q}_1 &= p_1, & \dot{p}_1 &= -q_1\lambda, \\ \dot{q}_2 &= p_2, & \dot{p}_2 &= -1 - q_2\lambda\end{aligned}\quad (4.5)$$

with $\lambda = (p_1^2 + p_2^2 - q_2)/(q_1^2 + q_2^2)$. This is a problem on the manifold

$$\mathcal{M} = \{(q_1, q_2, p_1, p_2) \mid q_1^2 + q_2^2 = 1, q_1 p_1 + q_2 p_2 = 0\}.$$

It is ρ -reversible with respect to $\rho(q_1, q_2, p_1, p_2) = (q_1, q_2, -p_1, -p_2)$ and also with respect to $\rho(q_1, q_2, p_1, p_2) = (-q_1, q_2, p_1, -p_2)$.

We apply two kinds of symmetric projection methods. First, we consider an orthogonal projection onto \mathcal{M} as in Algorithm 4.1. Second, we project parallel to coordinate axes. More precisely, we fix the first components in position and velocity if the angle of the pendulum is close to 0 or π (vertical projection in the picture to the right), and we fix the second components if the angle is close to $\pm\pi/2$ (horizontal projection). The regions where the direction of projection changes, are overlapping.



We notice in Fig. 4.3 that for the orthogonal projection method the energy error remains bounded, and this is also true for integrations over much longer time intervals. This is in agreement with the observation of Chap. I, where symmetric methods showed an excellent long-time behaviour when applied to reversible differential equations.

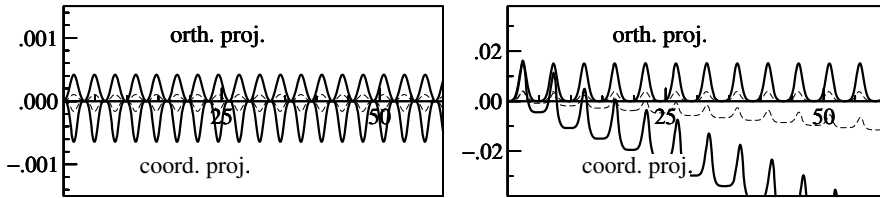


Fig. 4.3. Global error in the total energy for two different projection methods – orthogonal and coordinate projection – with the trapezoidal rule as basic integrator. Initial values for the position are $(\cos 0.8, -\sin 0.8)$ (left picture) and $(\cos 0.8, \sin 0.8)$ (right picture); zero initial values in the velocity; step sizes are $h = 0.1$ (solid) and $h = 0.05$ (thin dashed)

For the coordinate projection, however, we observe a bounded energy error only for the initial value that is close to equilibrium (no change in the direction of the projection is necessary). As soon as the direction has to be changed (right picture of Fig. 4.3) a linear drift in the energy error becomes visible. Hence, care has to be taken with the choice of the projection. For an explanation of this phenomenon we refer to Chap. IX on backward error analysis and to Chap. XI on perturbation theory of reversible mappings.

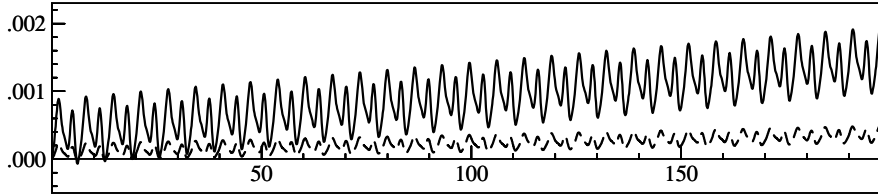


Fig. 4.4. Global error in the total energy for a symmetric projection method violating (1.4). Initial values for the position are $(\cos 0.8, -\sin 0.8)$ and $(0, 0)$ for the velocity; step sizes are $h = 0.1$ (solid) and $h = 0.05$ (thin dashed)

Example 4.4 (A Symmetric but Non-Reversible Projection Method). We consider the pendulum equation as in Example 4.3. This time, however, we apply a projection $\tilde{y}_n = y_n + A^T \mu$, $y_{n+1} = \tilde{y}_{n+1} + A^T \mu$ with

$$A = \begin{pmatrix} \varepsilon & 1 & 0 & 0 \\ \varepsilon & 0 & 0 & 1 \end{pmatrix}, \quad \varepsilon = 0.2.$$

For $\varepsilon = 0$ this corresponds to the vertical projection used in Example 4.3. For $\varepsilon \neq 0$ there is no matrix σ such that $\rho A^T = A^T \sigma$ holds for one of the mappings ρ that make the problem ρ -reversible. Hence condition (1.4) is violated, and the method is thus not ρ -reversible. The initial values are chosen such that $g'(y)A^T$ is invertible and well-conditioned along the solution. Although the projection direction need not be changed during the integration and the method is symmetric, the long-time behaviour is disappointing as shown in Fig. 4.4. This experiment illustrates that condition (1.4) is also important for a qualitatively correct simulation.

V.4.2 Symmetric Methods Based on Local Coordinates

Numerical methods for differential equations on manifolds that are based on local coordinates (Algorithm IV.5.3) are in general not symmetric. For example, if we consider the parametrization (IV.5.8) with respect to the tangent space at y_0 , the adjoint method would be parametrized by the tangent space at y_1 . We can circumvent this difficulty by the following algorithm (Hairer 2001).

Algorithm 4.5 (Symmetric Local Coordinates Approach). Assume that $y_n \in \mathcal{M}$ and that ψ_a is a local parametrization of \mathcal{M} satisfying $\psi_a(0) = a$ (close to y_n). One step $y_n \mapsto y_{n+1}$ is defined as follows (see Fig. 4.5):

- find z_n (close to 0) such that $\psi_a(z_n) = y_n$;
- $\tilde{z}_{n+1} = \Phi_h(z_n)$ (symmetric one-step method applied to (IV.5.7));
- $y_{n+1} = \psi_a(\tilde{z}_{n+1})$;
- choose a in the parametrization such that $z_n + \tilde{z}_{n+1} = 0$.

It is important to remark that the parametrization $y = \psi_a(z)$ is in general changed in every step.

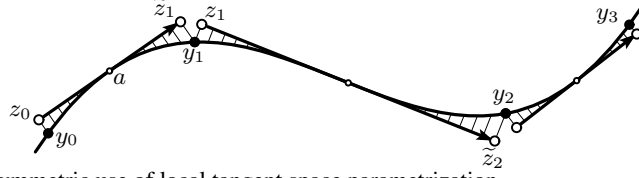


Fig. 4.5. Symmetric use of local tangent space parametrization

This algorithm is illustrated in Fig. 4.5 for the tangent space parametrization (IV.5.8), given by

$$\psi_a(z) = a + Q(a)z + g'(a)^T u_a(z), \quad (4.6)$$

where the columns of $Q(a)$ form an orthogonal basis of $T_a\mathcal{M}$ and the function $u_a(z)$ is such that $\psi_a(z) \in \mathcal{M}$. It satisfies $u_a(0) = 0$ and $u'_a(0) = 0$.

Existence of the Numerical Solution. In Algorithm 4.5 the values $a \in \mathcal{M}$ and z_n are implicitly determined by

$$F(h, z_n, a) = \begin{pmatrix} z_n + \Phi_h(z_n) \\ \psi_a(z_n) - y_n \end{pmatrix} = 0, \quad (4.7)$$

and the numerical solution is then explicitly given by $y_{n+1} = \psi_a(\Phi_h(z_n))$. For more clarity we also use here the notation $\psi(z, a) = \psi_a(z)$. If the parametrization $\psi(z, a)$ is differentiable, we have

$$\frac{\partial F}{\partial(z_n, a)}(0, 0, y_n) = \begin{pmatrix} 2I & 0 \\ \frac{\partial \psi}{\partial z}(0, y_n) & \frac{\partial \psi}{\partial a}(0, y_n) \end{pmatrix}. \quad (4.8)$$

Since $\psi(z, a) \in \mathcal{M}$ for all z and $a \in \mathcal{M}$, the derivative with respect to a lies in the tangent space. Assume now that the parametrization $\psi(z, a)$ is such that the restriction of $\frac{\partial \psi}{\partial a}(0, y_n)$ onto the tangent space $T_{y_n}\mathcal{M}$ is bijective. Then, the matrix (4.8) is invertible on $\mathbb{R}^d \times T_{y_n}\mathcal{M}$ (d denotes the dimension of the manifold). The implicit function theorem thus proves the existence of a numerical solution (z_n, a) close to $(0, y_n)$. In the case where $\psi_a(z)$ is given by (4.6), the matrix

$$\frac{\partial \psi}{\partial a}(0, a) = I - g'(a)^T (g'(a)g'(a)^T)^{-1} g'(a)$$

is a projection onto the tangent space $T_a\mathcal{M}$ and satisfies the above assumptions provided that $g'(a)$ has full rank.

Order. We let $y_n := y(t_n)$ be a value on the exact solution $y(t)$ of (4.1). Then we fix $a \in \mathcal{M}$ as follows: we replace the upper part of the definition (4.7) of $F(h, z_n, a)$ with $z_n + \varphi_h^{(z)}(z_n)$, where $\varphi_t^{(z)}$ denotes the exact flow of the differential equation for $z(t)$ equivalent to (4.1). The above considerations show that such an a exists; let us call it a^* . If Φ_h is of order p , we then have $F(h, z(t_n), a^*) = \mathcal{O}(h^{p+1})$. An application of the implicit function theorem thus gives $z_n - z(t_n) = \mathcal{O}(h^{p+1})$, implying $\tilde{z}_{n+1} - z(t_{n+1}) = \mathcal{O}(h^{p+1})$, and finally also $y_{n+1} - y(t_{n+1}) = \mathcal{O}(h^{p+1})$. This proves order p for the method defined by Algorithm 4.5.

Symmetry. Exchanging $h \leftrightarrow -h$ and $y_n \leftrightarrow y_{n+1}$ in Algorithm 4.5 yields

$$\psi_a(z_n) = y_{n+1}, \quad \tilde{z}_{n+1} = \Phi_{-h}(z_n), \quad y_n = \psi_a(\tilde{z}_{n+1}), \quad z_n + \tilde{z}_{n+1} = 0.$$

If we also exchange the auxiliary variables z_n and \tilde{z}_{n+1} and if we use the symmetry of the basic method Φ_h , we regain the original formulas. This proves the symmetry of the algorithm. Again various kinds of modifications are possible. For example, the condition $z_n + \tilde{z}_{n+1} = 0$ can be replaced with $z_n + \tilde{z}_{n+1} = \chi(h, z_n, \tilde{z}_{n+1})$. If $\chi(-h, v, u) = \chi(h, u, v)$, the symmetry of Algorithm 4.5 is not destroyed.

Reversibility. In general, we cannot expect the method of Algorithm 4.5 to satisfy the ρ -compatibility condition (1.4), which is needed for ρ -reversibility. However, if the parametrization is such that

$$\rho \psi_a(z) = \psi_{\rho a}(\sigma z) \quad \text{for some invertible } \sigma, \quad (4.9)$$

we shall show that the compatibility condition (1.4) holds. We first prove that for a ρ -reversible problem $\dot{y} = f(y)$ the differential equation (IV.5.7), written as $\dot{z} = F_a(z)$, is σ -reversible in the sense that $\sigma F_a(z) = -F_{\rho a}(\sigma z)$. This follows from $\rho \psi'_a(z) = \psi'_{\rho a}(\sigma z)\sigma$ (which is seen by differentiation of (4.9)) and from $f(\psi_{\rho a}(\sigma z)) = -\rho f(\psi_a(z))$, because

$$\psi'_a(z)F_a(z) = f(\psi_a(z)) \implies \psi'_{\rho a}(\sigma z)\sigma F_a(z) = -f(\psi_{\rho a}(\sigma z)).$$

If the basic method Φ_h satisfies $\sigma \circ \Phi_h = \Phi_{-h} \circ \sigma$ when applied to $\dot{z} = F_a(z)$ (e.g., for all Runge–Kutta methods), the formulas of Algorithm 4.5 satisfy

$$\begin{aligned} \rho y_n &= \rho \psi_a(z_n) = \psi_{\rho a}(\sigma z_n), & \sigma \tilde{z}_{n+1} &= \Phi_{-h}(\sigma z_n), \\ \psi_{\rho a}(\sigma \tilde{z}_{n+1}) &= \rho \psi_a(\tilde{z}_{n+1}) = \rho y_{n+1}, & \sigma z_n + \sigma \tilde{z}_{n+1} &= 0. \end{aligned}$$

This proves that, starting with ρy_n and a negative step size $-h$, the Algorithm 4.5 produces ρy_{n+1} , where y_{n+1} is just the result obtained with initial value y_n and step size h . But this is nothing other than the ρ -compatibility condition (1.4) for Algorithm 4.5.

In order to verify condition (4.9) for the tangent space parametrization (4.6), we write it as $\psi_a(Z) = a + Z + N(Z)$, where Z is an arbitrary element of the tangent space $T_a\mathcal{M}$ and $N(Z)$ is orthogonal to $T_a\mathcal{M}$ such that $\psi_a(Z) \in \mathcal{M}$. Since $\rho T_a\mathcal{M} = T_{\rho a}\mathcal{M}$ and since, for a ρ satisfying $\rho\rho^T = I$, the vector $\rho N(Z)$ is orthogonal to $T_{\rho a}\mathcal{M}$, we have $\rho\psi_a(Z) = \psi_{\rho a}(\rho Z)$. This proves (4.9) for the tangent space parametrization of a manifold.

Example 4.6. We repeated the experiment of Example 4.2 with Algorithm IV.5.3, using tangent space parametrization and the trapezoidal rule as basic integrator, and compared it to the symmetrized version of Algorithm 4.5. We were surprised to see that both algorithms worked equally well and gave a numerical solution lying near a closed curve. An explanation is given in Exercise 11. There it is shown that for the

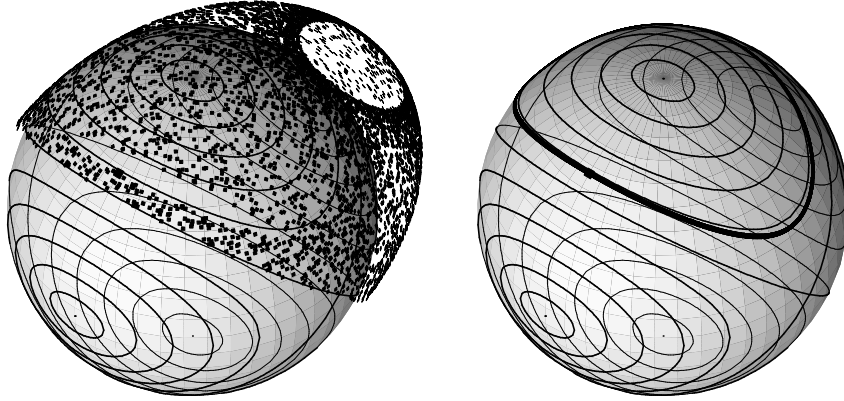


Fig. 4.6. Numerical simulation of the rigid body equations; standard use of tangent space parametrization with the trapezoidal rule as basic method (left picture) and its symmetrized version (right picture); 5000 steps with step size $h = 0.4$

special situation where \mathcal{M} is a sphere, the standard algorithm is also symmetric for the trapezoidal rule. Let us therefore modify the problem slightly.

We consider the rigid body equations (IV.1.4) as a differential equation on the manifold

$$\mathcal{M} = \left\{ (y_1, y_2, y_3) \mid \frac{y_1^2}{I_1} + \frac{y_2^2}{I_2} + \frac{y_3^2}{I_3} = \text{Const} \right\} \quad (4.10)$$

with parameters and initial data as in Example 4.2, and we apply the standard and the symmetrized method based on tangent space parametrization. The result is shown in Fig. 4.6. In both cases the numerical solution lies on the manifold (by definition of the method), but only the symmetric method has a correct long-time behaviour.

Symmetric Lie Group Methods. We turn our attention to particular problems

$$\dot{Y} = A(Y)Y, \quad Y(0) = Y_0, \quad (4.11)$$

where $A(Y)$ is in the Lie algebra \mathfrak{g} whenever Y is in the corresponding Lie group G . The exact solution then evolves on the manifold G . Munthe-Kaas methods (Sect. IV.8.2) are in general not symmetric, even if the underlying Runge–Kutta method is symmetric. This is due to the unsymmetric use of the local coordinates $Y = \exp(\Omega)Y_0$. However, accidentally, the Lie group method based on the implicit midpoint rule

$$Y_{n+1} = \exp(\Omega)Y_n, \quad \Omega = hA(\exp(\Omega/2)Y_n) \quad (4.12)$$

is symmetric. This can be seen as usual by exchanging $h \leftrightarrow -h$ and $Y_n \leftrightarrow Y_{n+1}$ (and also $\Omega \leftrightarrow -\Omega$ for the auxiliary variable). Numerical computations with the rigid body equations (considered as a problem on the sphere) shows an excellent long-time behaviour for this method similar to that of the right picture in Fig. 4.6. In contrast to the implicit midpoint rule (I.1.7), the numerical solution of (4.12) does not lie exactly on the ellipsoid (4.10); see Exercise 12.

For the construction of further symmetric Lie group methods we can apply the ideas of Algorithm 4.5. As local parametrization we choose

$$\psi_U(\Omega) = \exp(\Omega)U, \quad (4.13)$$

where $U = \exp(\Theta)Y_n$ plays the role of the midpoint on the manifold. We put $Z_n = -\Theta$ so that $\psi_U(Z_n) = Y_n$. With this starting value Z_n we apply any symmetric Runge–Kutta method to the differential equation

$$\dot{\Omega} = A(\psi_U(\Omega)) + \sum_{k=1}^q \frac{B_k}{k!} \text{ad}_{\Omega}^k \left(A(\psi_U(\Omega)) \right), \quad \Omega(0) = -\Theta, \quad (4.14)$$

(cf. (IV.8.9)) and thus obtain \tilde{Z}_{n+1} . According to Algorithm 4.5, Θ is implicitly determined by the condition $Z_n + \tilde{Z}_{n+1} = 0$, and the numerical approximation is obtained from

$$Y_{n+1} = \psi_U(\tilde{Z}_{n+1}) = \exp(\tilde{Z}_{n+1}) \exp(\Theta)Y_n = \exp(2\Theta)Y_n.$$

The method obtained in this way is identical to Algorithm 2 of Zanna, Engø & Munthe-Kaas (2001). With the coefficients of the 2-stage Gauss method (Table II.1.1) and with $q = 1$ in (4.14) we thus get

$$\begin{aligned} \Omega_1 &= -h \frac{\sqrt{3}}{6} \left(A_2 - \frac{1}{2} [\Omega_2, A_2] \right), & \Omega_2 &= h \frac{\sqrt{3}}{6} \left(A_1 - \frac{1}{2} [\Omega_1, A_1] \right) \\ Y_{n+1} &= \exp(2\Theta)Y_n = \exp \left(\frac{h}{2} (A_1 + A_2) - \frac{h}{4} ([\Omega_1, A_1] + [\Omega_2, A_2]) \right) Y_n, \end{aligned}$$

where $A_i = A(\exp(\Omega_i) \exp(\Theta)Y_n)$. This is a symmetric Lie group method of order four. We can reduce the number of commutators by replacing Ω_i in the right-hand expression with its dominating term. This yields

$$\begin{aligned} \Omega_1 &= -h \frac{\sqrt{3}}{6} A_2 + \frac{h^2}{24} [A_1, A_2], & \Omega_2 &= h \frac{\sqrt{3}}{6} A_1 - \frac{h^2}{24} [A_1, A_2] \\ Y_{n+1} &= \exp \left(\frac{h}{2} (A_1 + A_2) - h^2 \frac{\sqrt{3}}{12} [A_1, A_2] \right) Y_n \end{aligned} \quad (4.15)$$

(cf. Exercise IV.19). Although we have neglected terms of size $\mathcal{O}(h^4)$, the method remains of order four, because the order of symmetric methods is always even.

For any linear invertible transformation ρ , the parametrization (4.13) satisfies

$$\rho \psi_U(\Omega) = \rho \exp(\Omega)U = \exp(\rho \Omega \rho^{-1}) \rho U = \psi_{\rho U}(\sigma U)$$

with $\sigma \Omega = \rho \Omega \rho^{-1}$. Hence (4.9) holds true. If the problem (4.11) is ρ -reversible, i.e., $\rho A(Y) = -A(\rho Y)\rho$, then the truncated differential equation (4.14) is σ -reversible for all choices of the truncation index q . Moreover, after the simplifications that lead to method (4.15), the ρ -compatibility condition (1.4) is also satisfied.

The following variant is also proposed in Zanna, Engø & Munthe-Kaas (2001). Instead of computing Θ from the relation $Z_n + \tilde{Z}_{n+1} = 0$, Θ is determined by

$$Z_n + \tilde{Z}_{n+1} = h \sum_{i=1}^s e_i \left(A_i - \frac{1}{2} [\Omega_i, A_i] + \dots \right).$$

If the coefficients satisfy $e_{s+1-i} = -e_i$, this modification gives symmetric Lie group methods.

V.5 Energy – Momentum Methods and Discrete Gradients

Conventional numerical methods, when applied to the ordinary differential equations of motion of classical mechanics, conserve the total energy and angular momentum only to the order of the truncation error. Since these constants of motion play a central role in mechanics, it is a great advantage to be able to conserve them exactly.

(R.A. LaBudde & D. Greenspan 1976)

This section is concerned with numerical integrators for the equations of motion of classical mechanics which conserve both the total energy and angular momentum. Their construction is related to the concept of discrete gradients. The methods considered are symmetric, which is incidental but useful: in our view their good long-time behaviour is a consequence of their symmetry (and reversibility) more than of their exact conservation properties; see the disappointing behaviour of the non-symmetric energy- and momentum-conserving projection method in Example IV.4.4.

A Modified Midpoint Rule. Consider first a single particle of mass m in \mathbb{R}^3 , with position coordinates $q(t) \in \mathbb{R}^3$, moving in a central force field with potential $U(q) = V(\|q\|)$ (e.g., $V(r) = -1/r$ in the Kepler problem). With the momenta $p(t) = m \dot{q}(t)$, the equations of motion read

$$\dot{q} = \frac{1}{m} p, \quad \dot{p} = -\nabla U(q) = -V'(\|q\|) \frac{q}{\|q\|}.$$

Constants of motion are the total energy $H = T(p) + U(q)$, with $T(p) = \|p\|^2/(2m)$, and the angular momentum $L = q \times p$:

$$\frac{d}{dt}(q \times p) = \dot{q} \times p + q \times \dot{p} = \frac{1}{m} p \times p - V'(\|q\|) \frac{1}{\|q\|} q \times q = 0.$$

We know from Sect. IV.2 that the implicit midpoint rule conserves the quadratic invariant $L = q \times p$, and Theorem IV.2.4 (or a simple direct calculation) shows that L remains actually conserved by any modification of the form

$$\begin{aligned}
q_{n+1} &= q_n + \frac{h}{m} p_{n+1/2} & \text{with} & & p_{n+1/2} &= \frac{1}{2}(p_n + p_{n+1}) \\
p_{n+1} &= p_n - \kappa h \nabla U(q_{n+1/2}) & & & q_{n+1/2} &= \frac{1}{2}(q_n + q_{n+1})
\end{aligned} \tag{5.1}$$

where κ is an arbitrary real number. Simo, Tarnow & Wong (1992) introduce this additional parameter κ and determine it so that the total energy is conserved: $H(p_{n+1}, q_{n+1}) = H(p_n, q_n)$. With the notation $F_{n+1/2} = -\nabla U(q_{n+1/2}) = -V'(\|q_{n+1/2}\|)/\|q_{n+1/2}\| \cdot q_{n+1/2}$ we have

$$T(p_{n+1}) = T(p_n + \kappa h F_{n+1/2}) = T(p_n) + \frac{\kappa h}{m} p_{n+1/2}^T F_{n+1/2} ,$$

and hence the condition for conservation of the total energy $H = T + U$ becomes

$$\kappa \frac{h}{m} p_{n+1/2}^T F_{n+1/2} = U(q_n) - U(q_{n+1}) .$$

This gives a reasonable method even if $p_{n+1/2}^T F_{n+1/2}$ is arbitrarily close to zero. This is seen as follows: let $\sigma = -\kappa V'(\|q_{n+1/2}\|)/\|q_{n+1/2}\|$ so that $\kappa F_{n+1/2} = \sigma q_{n+1/2}$. The above condition for energy conservation then reads

$$\sigma \frac{h}{m} p_{n+1/2}^T q_{n+1/2} = V(\|q_n\|) - V(\|q_{n+1}\|) ,$$

where we note further that

$$\begin{aligned}
\frac{h}{m} p_{n+1/2}^T q_{n+1/2} &= (q_{n+1} - q_n)^T \frac{1}{2}(q_{n+1} + q_n) \\
&= \frac{1}{2}(\|q_{n+1}\|^2 - \|q_n\|^2) = (\|q_{n+1}\| - \|q_n\|) \frac{1}{2}(\|q_{n+1}\| + \|q_n\|) .
\end{aligned}$$

These formulas give

$$\sigma = - \frac{V(\|q_{n+1}\|) - V(\|q_n\|)}{\|q_{n+1}\| - \|q_n\|} \frac{1}{\frac{1}{2}(\|q_{n+1}\| + \|q_n\|)} , \tag{5.2}$$

with which method (5.1) becomes

$$\begin{aligned}
q_{n+1} &= q_n + \frac{h}{m} p_{n+1/2} \\
p_{n+1} &= p_n - h \frac{V(\|q_{n+1}\|) - V(\|q_n\|)}{\|q_{n+1}\| - \|q_n\|} \frac{q_{n+1/2}}{\frac{1}{2}(\|q_{n+1}\| + \|q_n\|)} .
\end{aligned} \tag{5.3}$$

This is a second-order symmetric method which conserves the total energy and the angular momentum. It evaluates only the potential $U(q) = V(\|q\|)$. The force $-\nabla U(q) = -V'(\|q\|) \frac{q}{\|q\|}$ is approximated by finite differences.

The energy- and momentum-conserving method (5.3) first appeared in LaBudde & Greenspan (1974). The method (5.1) or (5.3) is the starting point for extensions in several directions to other problems of mechanics and other methods; see Simo,

Tarnow & Wong (1992), Simo & Tarnow (1992), Lewis & Simo (1994, 1996), Gonzalez & Simo (1996), Gonzalez (1996), and Reich (1996b). In the following we consider a direct generalization to systems of particles, also given in LaBudde & Greenspan (1974).

An Energy-Momentum Method for N-Body Systems. We consider a system of N particles interacting pairwise with potential forces which depend on the distances between the particles. As in Example IV.1.3, this is formulated as a Hamiltonian system with total energy

$$H(p, q) = \frac{1}{2} \sum_{i=1}^N \frac{1}{m_i} p_i^T p_i + \sum_{i=2}^N \sum_{j=1}^{i-1} V_{ij}(\|q_i - q_j\|). \quad (5.4)$$

As an extension of method (5.3), we consider the following method (where we now write the time step number as a superscript for notational convenience)

$$\begin{aligned} q_i^{n+1} &= q_i^n + \frac{h}{m_i} p_i^{n+1/2} \\ p_i^{n+1} &= p_i^n + h \sum_{j=1}^N \sigma_{ij} (q_i^{n+1/2} - q_j^{n+1/2}) \end{aligned} \quad (5.5)$$

where $p_i^{n+1/2} = \frac{1}{2}(p_i^n + p_i^{n+1})$, $q_i^{n+1/2} = \frac{1}{2}(q_i^n + q_i^{n+1})$, and for $i > j$,

$$\sigma_{ij} = \sigma_{ji} = -\frac{V_{ij}(r_{ij}^{n+1}) - V_{ij}(r_{ij}^n)}{r_{ij}^{n+1} - r_{ij}^n} \frac{1}{\frac{1}{2}(r_{ij}^n + r_{ij}^{n+1})} \quad (5.6)$$

with $r_{ij}^n = \|q_i^n - q_j^n\|$, and $\sigma_{ii} = 0$. This method has the following properties.

Theorem 5.1 (LaBudde & Greenspan 1974). *The method (5.5) with (5.6) is a second-order symmetric implicit method which conserves the total linear momentum $P = \sum_{i=1}^N p_i$, the total angular momentum $L = \sum_{i=1}^N q_i \times p_i$, and the total energy H .*

Proof. A comparison of (5.6) with the equations of motion shows that the method is of order 2. Similar to the continuous case (Example IV.1.3), the conservation of linear and angular momentum is obtained as a consequence of the symmetry $\sigma_{ij} = \sigma_{ji}$ for all i, j . For the linear momentum we have

$$\sum_{i=1}^N p_i^{n+1} = \sum_{i=1}^N p_i^n + h \sum_{i=1}^N \sum_{j=1}^N \sigma_{ij} (q_i^{n+1/2} - q_j^{n+1/2}) = \sum_{i=1}^N p_i^n.$$

For the proof of the conservation of the angular momentum we observe that the first equation of (5.5) together with $p_i^{n+1/2} = \frac{1}{2}(p_i^{n+1} + p_i^n)$ yields

$$(q_i^{n+1} - q_i^n) \times (p_i^{n+1} + p_i^n) = 0 \quad (5.7)$$

for all i . The second equation of (5.5) together with $q_i^{n+1/2} = \frac{1}{2}(q_i^{n+1} + q_i^n)$ gives

$$\sum_{i=1}^N (q_i^{n+1} + q_i^n) \times (p_i^{n+1} - p_i^n) = 0, \quad (5.8)$$

because $\sigma_{ij} = \sigma_{ji}$ and therefore $\sum_{i,j=1}^N \sigma_{ij} q_i^{n+1/2} \times q_j^{n+1/2} = 0$. Adding the sum over i of (5.7) to the equation (5.8) proves the statement $\sum_{i=1}^N q_i^{n+1} \times p_i^{n+1} = \sum_{i=1}^N q_i^n \times p_i^n$.

It remains to show the energy conservation. Now, the kinetic energy $T(p) = \frac{1}{2} \sum_{i=1}^N m_i^{-1} p_i^T p_i$ at step $n+1$ is

$$\begin{aligned} T(p^{n+1}) &= T(p^n) + \sum_{i=1}^N \left(\frac{h}{m_i} p_i^{n+1/2} \right)^T \sum_{j=1}^N \sigma_{ij} (q_i^{n+1/2} - q_j^{n+1/2}) \\ &= T(p^n) + \sum_{i=1}^N \sum_{j=1}^N \sigma_{ij} (q_i^{n+1} - q_i^n)^T (q_i^{n+1/2} - q_j^{n+1/2}). \end{aligned}$$

Using once more the symmetry $\sigma_{ij} = \sigma_{ji}$, the double sum reduces to

$$\begin{aligned} &\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \sigma_{ij} \left((q_i^{n+1} - q_j^{n+1}) - (q_i^n - q_j^n) \right)^T \frac{1}{2} \left((q_i^{n+1} - q_j^{n+1}) + (q_i^n - q_j^n) \right) \\ &= \sum_{i=2}^N \sum_{j=1}^{i-1} \sigma_{ij} \frac{1}{2} \left((r_{ij}^{n+1})^2 - (r_{ij}^n)^2 \right). \end{aligned}$$

On the other hand, the change in the potential energy is

$$U(q^{n+1}) - U(q^n) = \sum_{i=2}^N \sum_{j=1}^{i-1} \left(V_{ij}(r_{ij}^{n+1}) - V_{ij}(r_{ij}^n) \right),$$

and hence (5.6) yields the conservation of the total energy $H = T + U$. \square

Discrete-Gradient Methods. The methods (5.3) and (5.5) are of the form

$$y_{n+1} = y_n + h \bar{B}(y_{n+1}, y_n) \bar{\nabla} H(y_{n+1}, y_n) \quad (5.9)$$

where $\bar{B}(\hat{y}, y)$ is a skew-symmetric matrix for all \hat{y}, y , and $\bar{\nabla} H(\hat{y}, y)$ is a *discrete gradient* of H , that is, a continuous function of (\hat{y}, y) satisfying

$$\begin{aligned} \bar{\nabla} H(\hat{y}, y)^T (\hat{y} - y) &= H(\hat{y}) - H(y) \\ \bar{\nabla} H(y, y) &= \nabla H(y). \end{aligned} \quad (5.10)$$

The symmetry of the methods is seen from the properties $\bar{B}(\hat{y}, y) = \bar{B}(y, \hat{y})$ and $\bar{\nabla} H(\hat{y}, y) = \bar{\nabla} H(y, \hat{y})$. For example, for method (5.3) we have, with $y = (p, q)$ and $\hat{y} = (\hat{p}, \hat{q})$,

$$\overline{B}(\hat{y}, y) = \begin{pmatrix} 0 & -I_3 \\ I_3 & 0 \end{pmatrix} \quad \text{and} \quad \overline{\nabla} H(\hat{y}, y) = \begin{pmatrix} \frac{1}{2}(\hat{p} + p) \\ \sigma(\hat{q}, q) \frac{1}{2}(\hat{q} + q) \end{pmatrix}$$

where $\sigma(\hat{q}, q)$ is given by the expression (5.2) with (\hat{q}, q) in place of (q_{n+1}, q_n) or by the corresponding limit as $\|\hat{q}\| \rightarrow \|q\|$.

The discrete-gradient method (5.9) is consistent with the differential equation

$$\dot{y} = B(y) \nabla H(y) \quad (5.11)$$

with the skew-symmetric matrix $B(y) = \overline{B}(y, y)$. This system conserves H , since

$$\frac{d}{dt} H(y) = \nabla H(y)^T \dot{y} = \nabla H(y)^T B(y) \nabla H(y) = 0,$$

and, as was noted by Gonzalez (1996) and McLachlan, Quispel & Robidoux (1999), H is also conserved by method (5.9).

Theorem 5.2. *The discrete-gradient method (5.9) conserves the invariant H of the system (5.11).*

Proof. The definitions (5.10) of a discrete gradient and of the method (5.9) give

$$\begin{aligned} H(y_{n+1}) - H(y_n) &= \overline{\nabla} H(y_{n+1}, y_n)^T (y_{n+1} - y_n) \\ &= h \overline{\nabla} H(y_{n+1}, y_n)^T \overline{B}(y_{n+1}, y_n) \overline{\nabla} H(y_{n+1}, y_n) = 0, \end{aligned}$$

where the last equality follows from the skew-symmetry of $\overline{B}(y_{n+1}, y_n)$. \square

Example 5.3. The Lotka–Volterra system (I.1.1) can be written as

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} 0 & -uv \\ uv & 0 \end{pmatrix} \nabla H(u, v)$$

with the invariant $H(u, v) = \ln u - u + 2 \ln v - v$ of (I.1.4). Possible choices of a discrete gradient are the *coordinate increment discrete gradient* (Itoh & Abe 1988)

$$\overline{\nabla} H(\hat{u}, \hat{v}; u, v) = \begin{pmatrix} \frac{H(\hat{u}, v) - H(u, v)}{\hat{u} - u} \\ \frac{H(\hat{u}, \hat{v}) - H(\hat{u}, v)}{\hat{v} - v} \end{pmatrix} \quad (5.12)$$

and the *midpoint discrete gradient* (Gonzalez 1996)

$$\overline{\nabla} H(\hat{y}, y) = \nabla H(\bar{y}) + \frac{H(\hat{y}) - H(y) - \nabla H(\bar{y})^T \Delta y}{\|\Delta y\|^2} \Delta y \quad (5.13)$$

with $\bar{y} = \frac{1}{2}(\hat{y} + y)$ and $\Delta y = \hat{y} - y$. In contrast to (5.12), the discrete gradient (5.13) yields a symmetric discretization.

A systematic study of discrete-gradient methods is given in Gonzalez (1996) and McLachlan, Quispel & Robidoux (1999).

V.6 Exercises

1. Prove that (after a suitable permutation of the stages) the condition $c_{s+1-i} = 1 - c_i$ (for all i) is also necessary for a collocation method to be symmetric.
2. Prove that explicit Runge–Kutta methods cannot be symmetric.
Hint. If a one-step method applied to $\dot{y} = \lambda y$ yields $y_1 = R(h\lambda)y_0$ then, a necessary condition for the symmetry of the method is $R(z)R(-z) = 1$ for all complex z .
3. Consider an irreducible diagonally implicit Runge–Kutta method (irreducible in the sense of Sect. VI.7.3). Prove that the condition (2.4) is necessary for the symmetry of the method. No permutation of the stages has to be performed.
4. Let $\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]}$, where $\varphi_t^{[i]}$ represents the exact flow of $\dot{y} = f^{[i]}(y)$. In the situation of Theorem III.3.17 prove that the local error (3.4) of the composition method (3.3) has the form

$$h^3 \left(\frac{1}{24} (6\alpha - 1) [D_2, [D_2, D_1]] + \frac{1}{12} (1 - 6\alpha + 6\alpha^2) [D_1, [D_1, D_2]] \right) Id(y),$$

where, as usual, $D_i g(y) = g'(y) f^{[i]}(y)$. The value $\alpha = 0.1932$ is found by minimizing the expression $(6\alpha - 1)^2 + 4(1 - 6\alpha + 6\alpha^2)^2$ (McLachlan 1995).

5. For the linear transformation $\rho(p, q) = (-p, q)$, consider a ρ -reversible problem (1.3) with scalar p and q . Prove that every solution which crosses the q -axis twice is periodic.
6. Prove that if a numerical method conserves quadratic invariants (IV.2.1), then so does its adjoint.
7. For the numerical solution of $\dot{y} = A(t)y$ consider the method $y_n \mapsto y_{n+1}$ defined by $y_{n+1} = z(t_n + h)$, where $z(t)$ is the solution of

$$\dot{z} = \hat{A}(t)z, \quad z(t_n) = y_n,$$

and $\hat{A}(t)$ is the interpolation polynomial based on symmetric nodes c_1, \dots, c_s , i.e., $c_{s+1-i} + c_i = 1$ for all i .

- a) Prove that this method is symmetric.
 - b) Show that $y_{n+1} = \exp(\Omega(h))y_n$ holds, where $\Omega(h)$ has an expansion in odd powers of h . This justifies the omission of the terms involving triple integrals in Example IV.7.4.
8. If Φ_h stands for the implicit midpoint rule, what are the Runge–Kutta coefficients of the composition method (3.8)? The general theory of Sect. III.1 gives three order conditions for order 4 (those for the trees of order 2 and 4 are automatically satisfied by the symmetry of the method). Are they compatible with the two conditions of Example 3.5?
 9. Make a numerical comparison of our favourite composition methods *p6 s9*, *p8 s17*, and *p10 s35* for the Lorenz problem

$$\begin{aligned} y_1' &= -\sigma(y_1 - y_2) & y_1(0) &= 10 & \sigma &= 10 \\ y_2' &= -y_1 y_3 + r y_1 - y_2 & y_2(0) &= -20 & r &= 28 \\ y_3' &= y_1 y_2 - b y_3 & y_3(0) &= 20 & b &= 8/3 \end{aligned} \quad (6.1)$$

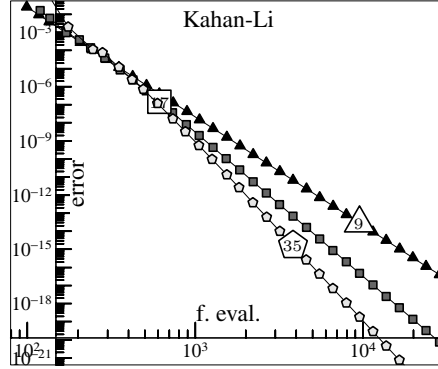


Fig. 6.1. Comparison of various composition methods applied to the Lorenz equations

with exact solution

$$\begin{aligned} y_1(1) &= 8.635692709892506017930544628639 \\ y_2(1) &= 2.798663387927457052023080059065 \\ y_3(1) &= 33.36063508973142157789185846267 \end{aligned} \quad (6.2)$$

by composing for $0 \leq t \leq 1$ the second order *symmetric splitting* scheme (see Kahan & Li 1997) which, for the time-stepping $y_i \mapsto Y_i$, is given by

$$\begin{aligned} Y_1 - y_1 &= \frac{h}{2}(-\sigma(y_1 + Y_1 - y_2 - Y_2)) \\ Y_2 - y_2 &= \frac{h}{2}(-y_1 Y_3 - Y_1 y_3 + r y_1 + r Y_1 - y_2 - Y_2) \\ Y_3 - y_3 &= \frac{h}{2}(y_1 Y_2 + Y_1 y_2 - b y_3 - b Y_3). \end{aligned} \quad (6.3)$$

This method requires, for each step, the solution of a *linear* system only. The results are shown in Fig. 6.1.

10. *Symmetrized order conditions* (Suzuki 1992). Prove that for methods (3.8) of order four with γ_i satisfying (3.10)

$$\sum_{k=1}^s \gamma_k^3 \left(\sum_{\ell=1}^k \gamma_\ell \right)^2 = 0 \quad \Longleftrightarrow \quad \sum_{k=1}^s \gamma_k^3 \left(\sum_{\ell=1}^k \gamma_\ell \right) \left(\sum_{\ell=k}^s \gamma_\ell \right) = 0.$$

The prime after (before) a sum sign indicates that the term with highest (lowest) index is divided by 2. Prove also that the order conditions given in Suzuki (1992) for order $p \leq 8$ are equivalent to those of Example 3.5. Is this also true for order $p = 10$?

Hint. Use relations like $\sum_{\ell=1}^k \gamma_\ell = 1 - \sum_{\ell=k}^s \gamma_\ell$.

11. Let $\mathcal{M} = \{(y_1, y_2, y_3) \mid y_1^2 + y_2^2 + y_3^2 = 1\}$, and consider for $a \in \mathcal{M}$ the tangent space parametrization

$$\psi_a(z) = a + z + a u_a(z),$$

where, for $z \in T_a\mathcal{M}$, the real value $u_a(z)$ is determined by the requirement $\psi_a(z) \in \mathcal{M}$. Prove that Algorithm IV.5.3, with the trapezoidal rule in the role of Φ_h , is a symmetric method.

Hint. Since z is a linear combination of a and $\psi_a(z)$, it is uniquely determined by $a^T z$ (which is zero) and $\psi_a(z)^T z$.

12. (Zanna, Engø & Munthe-Kaas 2001). Verify numerically that the Lie group method (4.12) based on the implicit midpoint rule does not conserve general quadratic first integrals. One can consider the rigid body equations in the form (IV.1.5).

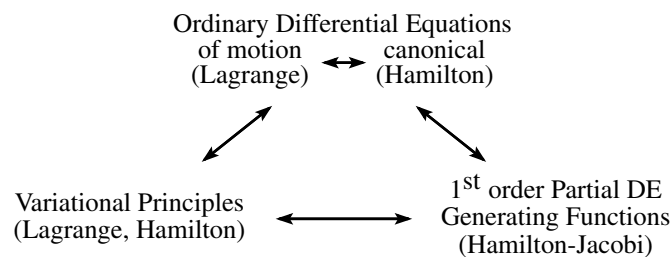
Chapter VI.

Symplectic Integration of Hamiltonian Systems



Fig. 0.1. Sir William Rowan Hamilton, born: 4 August 1805 in Dublin, died: 2 September 1865. Famous for research in optics, mechanics, and for the invention of quaternions

Hamiltonian systems form the most important class of ordinary differential equations in the context of ‘Geometric Numerical Integration’. An outstanding property of these systems is the symplecticity of the flow. As indicated in the following diagram,



Hamiltonian theory operates in three different domains (equations of motion, partial differential equations and variational principles) which are all interconnected. Each of these viewpoints, which we will study one after the other, leads to the construction of methods preserving the symplecticity.

VI.1 Hamiltonian Systems

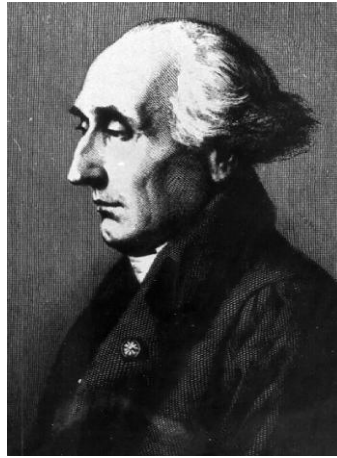
Hamilton's equations appeared first, among thousands of other formulas, and inspired by previous research in optics, in Hamilton (1834). Their importance was immediately recognized by Jacobi, who stressed and extended the fundamental ideas, so that, a couple of years later, all the long history of research of Galilei, Newton, Euler and Lagrange, was, in the words of Jacobi (1842), "to be considered as an introduction". The next mile-stones in the exposition of the theory were the monumental three volumes of Poincaré (1892, 1893, 1899) on celestial mechanics, Siegel's "Lectures on Celestial Mechanics" (1956), English enlarged edition by Siegel & Moser (1971), and the influential book of V.I. Arnold (1989; first Russian edition 1974). Beyond that, Hamiltonian systems became fundamental in many branches of physics. One such area, the dynamics of particle accelerators, actually motivated the construction of the first symplectic integrators (Ruth 1983).

VI.1.1 Lagrange's Equations

Équations différentielles pour la solution de tous les problèmes de Dynamique.
(J.-L. Lagrange 1788)

The problem of computing the dynamics of general mechanical systems began with Galilei (published 1638) and Newton's *Principia* (1687). The latter allowed one to reduce the movement of free mass points (the "mass points" being such planets as Mars or Jupiter) to the solution of differential equations (see Sect. I.2). But the movement of more complicated systems such as rigid bodies or bodies attached to each other by rods or springs, were the subject of long and difficult developments, until Lagrange (1760, 1788) found an elegant way of treating such problems in general.

We suppose that the position of a mechanical system with d degrees of freedom is described by $q = (q_1, \dots, q_d)^T$ as *generalized coordinates* (this can be for example Cartesian coordinates, angles, arc lengths along a curve, etc.). The theory is then built upon two pillars, namely an expression



Joseph-Louis Lagrange¹

$$T = T(q, \dot{q}) \tag{1.1}$$

which represents the *kinetic energy* (and which is often of the form $\frac{1}{2}\dot{q}^T M(q)\dot{q}$ where $M(q)$ is symmetric and positive definite), and by a function

¹ Joseph-Louis Lagrange, born: 25 January 1736 in Turin, Sardinia-Piedmont (now Italy), died: 10 April 1813 in Paris.

$$U = U(q) \quad (1.2)$$

representing the *potential energy*. Then, after denoting by

$$L = T - U \quad (1.3)$$

the corresponding *Lagrangian*, the coordinates $q_1(t), \dots, q_d(t)$ obey the differential equations

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}} \right) = \frac{\partial L}{\partial q}, \quad (1.4)$$

which constitute the *Lagrange equations* of the system. A numerical (or analytical) integration of these equations allows one to predict the motion of any such system from given initial values (“Ce sont ces équations qui serviront à déterminer la courbe décrite par le corps M et sa vitesse à chaque instant”; Lagrange 1760, p. 369).

Example 1.1. For a mass point of mass m in \mathbb{R}^3 with Cartesian coordinates $x = (x_1, x_2, x_3)^T$ we have $T(\dot{x}) = m(\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2)/2$. We suppose the point to move in a conservative force field $F(x) = -\nabla U(x)$. Then, the Lagrange equations (1.4) become $m\ddot{x} = F(x)$, which is Newton’s second law. The equations (1.2.2) for the planetary motion are precisely of this form.

Example 1.2 (Pendulum). For the mathematical pendulum of Sect. I.1 we take the angle α as coordinate. The kinetic and potential energies are given by $T = m(\dot{x}^2 + \dot{y}^2)/2 = m\ell^2\dot{\alpha}^2/2$ and $U = mgy = -mg\ell \cos \alpha$, respectively, so that the Lagrange equations become $-mg\ell \sin \alpha - m\ell^2\ddot{\alpha} = 0$ or equivalently $\ddot{\alpha} + \frac{g}{\ell} \sin \alpha = 0$.

VI.1.2 Hamilton’s Canonical Equations

An diese *Hamiltonsche* Form der Differentialgleichungen werden die ferneren Untersuchungen, welche den Kern dieser Vorlesung bilden, anknüpfen; das Bisherige ist als Einleitung dazu anzusehen.

(C.G.J. Jacobi 1842, p. 143)

Hamilton (1834) simplified the structure of Lagrange’s equations and turned them into a form that has remarkable symmetry, by

- introducing Poisson’s variables, the conjugate *momenta*

$$p_k = \frac{\partial L}{\partial \dot{q}_k}(q, \dot{q}) \quad \text{for } k = 1, \dots, d, \quad (1.5)$$

- considering the *Hamiltonian*

$$H := p^T \dot{q} - L(q, \dot{q}) \quad (1.6)$$

as a function of p and q , i.e., taking $H = H(p, q)$ obtained by expressing \dot{q} as a function of p and q via (1.5).

Here it is, of course, required that (1.5) defines, for every q , a continuously differentiable bijection $\dot{q} \leftrightarrow p$. This map is called the *Legendre transform*.

Theorem 1.3. *Lagrange's equations (1.4) are equivalent to Hamilton's equations*

$$\dot{p}_k = -\frac{\partial H}{\partial q_k}(p, q), \quad \dot{q}_k = \frac{\partial H}{\partial p_k}(p, q), \quad k = 1, \dots, d. \quad (1.7)$$

Proof. The definitions (1.5) and (1.6) for the momenta p and for the Hamiltonian H imply that

$$\begin{aligned} \frac{\partial H}{\partial p} &= \dot{q}^T + p^T \frac{\partial \dot{q}}{\partial p} - \frac{\partial L}{\partial \dot{q}} \frac{\partial \dot{q}}{\partial p} = \dot{q}^T, \\ \frac{\partial H}{\partial q} &= p^T \frac{\partial \dot{q}}{\partial q} - \frac{\partial L}{\partial q} - \frac{\partial L}{\partial \dot{q}} \frac{\partial \dot{q}}{\partial q} = -\frac{\partial L}{\partial q}. \end{aligned}$$

The Lagrange equations (1.4) are therefore equivalent to (1.7). \square

Case of Quadratic T . In the case that $T = \frac{1}{2}\dot{q}^T M(q)\dot{q}$ is quadratic, where $M(q)$ is a symmetric and positive definite matrix, we have, for a fixed q , $p = M(q)\dot{q}$, so that the existence of the Legendre transform is established. Further, by replacing the variable \dot{q} by $M(q)^{-1}p$ in the definition (1.6) of $H(p, q)$, we obtain

$$\begin{aligned} H(p, q) &= p^T M(q)^{-1}p - L(q, M(q)^{-1}p) \\ &= p^T M(q)^{-1}p - \frac{1}{2} p^T M(q)^{-1}p + U(q) = \frac{1}{2} p^T M(q)^{-1}p + U(q) \end{aligned}$$

and the Hamiltonian is $H = T + U$, which is the *total energy* of the mechanical system.

In Chap. I we have seen several examples of Hamiltonian systems, e.g., the pendulum (I.1.13), the Kepler problem (I.2.2), the outer solar system (I.2.12), etc. In the following we consider Hamiltonian systems (1.7) where the Hamiltonian $H(p, q)$ is arbitrary, and so not necessarily related to a mechanical problem.

VI.2 Symplectic Transformations

The name “complex group” formerly advocated by me in allusion to line complexes, ... has become more and more embarrassing through collision with the word “complex” in the connotation of complex number. I therefore propose to replace it by the Greek adjective “symplectic.”
(H. Weyl (1939), p. 165)

A first property of Hamiltonian systems, already seen in Example 1.2 of Sect. IV.1, is that the Hamiltonian $H(p, q)$ is a *first integral* of the system (1.7). In this section we shall study another important property – the *symplecticity* of its flow. The basic objects to be studied are two-dimensional parallelograms lying in \mathbb{R}^{2d} . We suppose the parallelogram to be spanned by two vectors

$$\xi = \begin{pmatrix} \xi^p \\ \xi^q \end{pmatrix}, \quad \eta = \begin{pmatrix} \eta^p \\ \eta^q \end{pmatrix}$$

in the (p, q) space ($\xi^p, \xi^q, \eta^p, \eta^q$ are in \mathbb{R}^d) as

$$P = \{t\xi + s\eta \mid 0 \leq t \leq 1, 0 \leq s \leq 1\}.$$

In the case $d = 1$ we consider the *oriented area*

$$\text{or.area}(P) = \det \begin{pmatrix} \xi^p & \eta^p \\ \xi^q & \eta^q \end{pmatrix} = \xi^p \eta^q - \xi^q \eta^p \quad (2.1)$$

(see left picture of Fig. 2.1). In higher dimensions, we replace this by the *sum of the oriented areas of the projections of P onto the coordinate planes (p_i, q_i)* , i.e., by

$$\omega(\xi, \eta) := \sum_{i=1}^d \det \begin{pmatrix} \xi_i^p & \eta_i^p \\ \xi_i^q & \eta_i^q \end{pmatrix} = \sum_{i=1}^d (\xi_i^p \eta_i^q - \xi_i^q \eta_i^p). \quad (2.2)$$

This defines a bilinear map acting on vectors of \mathbb{R}^{2d} , which will play a central role for Hamiltonian systems. In matrix notation, this map has the form

$$\omega(\xi, \eta) = \xi^T J \eta \quad \text{with} \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \quad (2.3)$$

where I is the identity matrix of dimension d .

Definition 2.1. A linear mapping $A : \mathbb{R}^{2d} \rightarrow \mathbb{R}^{2d}$ is called *symplectic* if

$$A^T J A = J$$

or, equivalently, if $\omega(A\xi, A\eta) = \omega(\xi, \eta)$ for all $\xi, \eta \in \mathbb{R}^{2d}$.

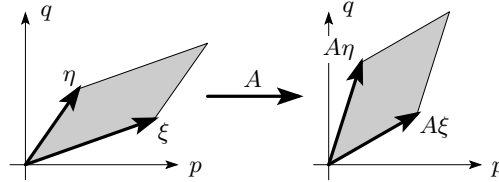


Fig. 2.1. Symplecticity (area preservation) of a linear mapping

In the case $d = 1$, where the expression $\omega(\xi, \eta)$ represents the area of the parallelogram P , symplecticity of a linear mapping A is therefore the *area preservation* of A (see Fig. 2.1). In the general case ($d > 1$), symplecticity means that the sum of the oriented areas of the projections of P onto (p_i, q_i) is the same as that for the transformed parallelograms $A(P)$.

We now turn our attention to nonlinear mappings. Differentiable functions can be locally approximated by linear mappings. This justifies the following definition.

Definition 2.2. A differentiable map $g : U \rightarrow \mathbb{R}^{2d}$ (where $U \subset \mathbb{R}^{2d}$ is an open set) is called *symplectic* if the Jacobian matrix $g'(p, q)$ is everywhere symplectic, i.e., if

$$g'(p, q)^T J g'(p, q) = J \quad \text{or} \quad \omega(g'(p, q)\xi, g'(p, q)\eta) = \omega(\xi, \eta).$$

Let us give a geometric interpretation of symplecticity for nonlinear mappings. Consider a 2-dimensional sub-manifold M of the $2d$ -dimensional set U , and suppose that it is given as the image $M = \psi(K)$ of a compact set $K \subset \mathbb{R}^2$, where

$\psi(s, t)$ is a continuously differentiable function. The manifold M can then be considered as the limit of a union of small parallelograms spanned by the vectors

$$\frac{\partial \psi}{\partial s}(s, t) ds \quad \text{and} \quad \frac{\partial \psi}{\partial t}(s, t) dt.$$

For one such parallelogram we consider (as above) the sum over the oriented areas of its projections onto the (p_i, q_i) plane. We then sum over all parallelograms of the manifold. In the limit this gives the expression

$$\Omega(M) = \iint_K \omega \left(\frac{\partial \psi}{\partial s}(s, t), \frac{\partial \psi}{\partial t}(s, t) \right) ds dt. \quad (2.4)$$

The transformation formula for double integrals implies that $\Omega(M)$ is independent of the parametrization ψ of M .

Lemma 2.3. *If the mapping $g : U \rightarrow \mathbb{R}^{2d}$ is symplectic on U , then it preserves the expression $\Omega(M)$, i.e.,*

$$\Omega(g(M)) = \Omega(M)$$

holds for all 2-dimensional manifolds M that can be represented as the image of a continuously differentiable function ψ .

Proof. The manifold $g(M)$ can be parametrized by $g \circ \psi$. We have

$$\Omega(g(M)) = \iint_K \omega \left(\frac{\partial(g \circ \psi)}{\partial s}(s, t), \frac{\partial(g \circ \psi)}{\partial t}(s, t) \right) ds dt = \Omega(M),$$

because $(g \circ \psi)'(s, t) = g'(\psi(s, t))\psi'(s, t)$ and g is a symplectic transformation. \square

For $d = 1$, M is already a subset of \mathbb{R}^2 and we choose $K = M$ with ψ the identity map. In this case, $\Omega(M) = \iint_M ds dt$ represents the area of M . Hence, Lemma 2.3 states that all symplectic mappings (also nonlinear ones) are *area preserving*.

We are now able to prove the main result of this section. We use the notation $y = (p, q)$, and we write the Hamiltonian system (1.7) in the form

$$\dot{y} = J^{-1} \nabla H(y), \quad (2.5)$$

where J is the matrix of (2.3) and $\nabla H(y) = H'(y)^T$.

Recall that the flow $\varphi_t : U \rightarrow \mathbb{R}^{2d}$ of a Hamiltonian system is the mapping that advances the solution by time t , i.e., $\varphi_t(p_0, q_0) = (p(t, p_0, q_0), q(t, p_0, q_0))$, where $p(t, p_0, q_0)$, $q(t, p_0, q_0)$ is the solution of the system corresponding to initial values $p(0) = p_0$, $q(0) = q_0$.

Theorem 2.4 (Poincaré 1899). *Let $H(p, q)$ be a twice continuously differentiable function on $U \subset \mathbb{R}^{2d}$. Then, for each fixed t , the flow φ_t is a symplectic transformation wherever it is defined.*

Proof. The derivative $\partial \varphi_t / \partial y_0$ (with $y_0 = (p_0, q_0)$) is a solution of the variational equation which, for the Hamiltonian system (2.5), is of the form $\dot{\Psi} = J^{-1} \nabla^2 H(\varphi_t(y_0)) \Psi$, where $\nabla^2 H(p, q)$ is the Hessian matrix of $H(p, q)$ ($\nabla^2 H(p, q)$

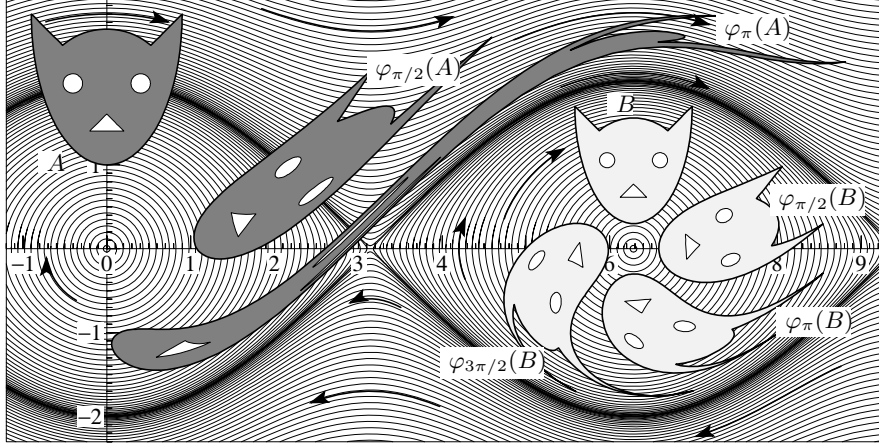


Fig. 2.2. Area preservation of the flow of Hamiltonian systems

is symmetric). We therefore obtain

$$\begin{aligned} \frac{d}{dt} \left(\left(\frac{\partial \varphi_t}{\partial y_0} \right)^T J \left(\frac{\partial \varphi_t}{\partial y_0} \right) \right) &= \left(\frac{d}{dt} \frac{\partial \varphi_t}{\partial y_0} \right)^T J \left(\frac{\partial \varphi_t}{\partial y_0} \right) + \left(\frac{\partial \varphi_t}{\partial y_0} \right)^T J \left(\frac{d}{dt} \frac{\partial \varphi_t}{\partial y_0} \right) \\ &= \left(\frac{\partial \varphi_t}{\partial y_0} \right)^T \nabla^2 H(\varphi_t(y_0)) J^{-T} J \left(\frac{\partial \varphi_t}{\partial y_0} \right) + \left(\frac{\partial \varphi_t}{\partial y_0} \right)^T \nabla^2 H(\varphi_t(y_0)) \left(\frac{\partial \varphi_t}{\partial y_0} \right) = 0, \end{aligned}$$

because $J^T = -J$ and $J^{-T} J = -I$. Since the relation

$$\left(\frac{\partial \varphi_t}{\partial y_0} \right)^T J \left(\frac{\partial \varphi_t}{\partial y_0} \right) = J \quad (2.6)$$

is satisfied for $t = 0$ (φ_0 is the identity map), it is satisfied for all t and all (p_0, q_0) , as long as the solution remains in the domain of definition of H . \square

Example 2.5. We illustrate this theorem with the pendulum problem (Example 1.2) using the normalization $m = \ell = g = 1$. We have $q = \alpha$, $p = \dot{\alpha}$, and the Hamiltonian is given by

$$H(p, q) = p^2/2 - \cos q.$$

Fig. 2.2 shows level curves of this function, and it also illustrates the area preservation of the flow φ_t . Indeed, by Theorem 2.4 and Lemma 2.3, the areas of A and $\varphi_t(A)$ as well as those of B and $\varphi_t(B)$ are the same, although their appearance is completely different.

We next show that symplecticity of the flow is a characteristic property for Hamiltonian systems. We call a differential equation $\dot{y} = f(y)$ *locally Hamiltonian*, if for every $y_0 \in U$ there exists a neighbourhood where $f(y) = J^{-1} \nabla H(y)$ for some function H .

Theorem 2.6. *Let $f : U \rightarrow \mathbb{R}^{2d}$ be continuously differentiable. Then, $\dot{y} = f(y)$ is locally Hamiltonian if and only if its flow $\varphi_t(y)$ is symplectic for all $y \in U$ and for all sufficiently small t .*

Proof. The necessity follows from Theorem 2.4. We therefore assume that the flow φ_t is symplectic, and we have to prove the local existence of a function $H(y)$ such that $f(y) = J^{-1}\nabla H(y)$. Differentiating (2.6) and using the fact that $\partial\varphi_t/\partial y_0$ is a solution of the variational equation $\dot{\Psi} = f'(\varphi_t(y_0))\Psi$, we obtain

$$\frac{d}{dt} \left(\left(\frac{\partial\varphi_t}{\partial y_0} \right)^T J \left(\frac{\partial\varphi_t}{\partial y_0} \right) \right) = \left(\frac{\partial\varphi_t}{\partial y_0} \right) \left(f'(\varphi_t(y_0))^T J + J f'(\varphi_t(y_0)) \right) \left(\frac{\partial\varphi_t}{\partial y_0} \right) = 0.$$

Putting $t = 0$, it follows from $J = -J^T$ that $Jf'(y_0)$ is a symmetric matrix for all y_0 . The Integrability Lemma 2.7 below shows that $Jf(y)$ can be written as the gradient of a function $H(y)$. \square

The following integrability condition for the existence of a potential was already known to Euler and Lagrange (see e.g., Euler's *Opera Omnia*, vol. 19. p. 2-3, or Lagrange (1760), p. 375).

Lemma 2.7 (Integrability Lemma). *Let $D \subset \mathbb{R}^n$ be open and $f : D \rightarrow \mathbb{R}^n$ be continuously differentiable, and assume that the Jacobian $f'(y)$ is symmetric for all $y \in D$. Then, for every $y_0 \in D$ there exists a neighbourhood and a function $H(y)$ such that*

$$f(y) = \nabla H(y) \quad (2.7)$$

on this neighbourhood. In other words, the differential form $f_1(y) dy_1 + \dots + f_n(y) dy_n = dH$ is a total differential.

Proof. Assume $y_0 = 0$, and consider a ball around y_0 which is contained in D . On this ball we define

$$H(y) = \int_0^1 y^T f(ty) dt + \text{Const.}$$

Differentiation with respect to y_k , and using the symmetry assumption $\partial f_i/\partial y_k = \partial f_k/\partial y_i$ yields

$$\frac{\partial H}{\partial y_k}(y) = \int_0^1 \left(f_k(ty) + y^T \frac{\partial f}{\partial y_k}(ty)t \right) dt = \int_0^1 \frac{d}{dt} (t f_k(ty)) dt = f_k(y),$$

which proves the statement. \square

For $D = \mathbb{R}^{2d}$ or for star-shaped regions D , the above proof shows that the function H of Lemma 2.7 is globally defined. Hence the Hamiltonian of Theorem 2.6 is also globally defined in this case. This remains valid for simply connected sets D . A counter-example, which shows that the existence of a global Hamiltonian in Theorem 2.6 is not true for general D , is given in Exercise 6.

An important property of symplectic transformations, which goes back to Jacobi (1836, "Theorem X"), is that they preserve the Hamiltonian character of the differential equation. Such transformations have been termed *canonical* since the 19th century. The next theorem shows that canonical and symplectic transformations are the same.

Theorem 2.8. *Let $\psi : U \rightarrow V$ be a change of coordinates such that ψ and ψ^{-1} are continuously differentiable functions. If ψ is symplectic, the Hamiltonian system $\dot{y} = J^{-1}\nabla H(y)$ becomes in the new variables $z = \psi(y)$*

$$\dot{z} = J^{-1}\nabla K(z) \quad \text{with} \quad K(z) = H(y). \quad (2.8)$$

Conversely, if ψ transforms every Hamiltonian system to another Hamiltonian system via (2.8), then ψ is symplectic.

Proof. Since $\dot{z} = \psi'(y)\dot{y}$ and $\psi'(y)^T \nabla K(z) = \nabla H(y)$, the Hamiltonian system $\dot{y} = J^{-1}\nabla H(y)$ becomes

$$\dot{z} = \psi'(y)J^{-1}\psi'(y)^T \nabla K(z) \quad (2.9)$$

in the new variables. It is equivalent to (2.8) if

$$\psi'(y)J^{-1}\psi'(y)^T = J^{-1}. \quad (2.10)$$

Multiplying this relation from the right by $\psi'(y)^{-T}$ and from the left by $\psi'(y)^{-1}$ and then taking its inverse yields $J = \psi'(y)^T J \psi'(y)$, which shows that (2.10) is equivalent to the symplecticity of ψ .

For the inverse relation we note that (2.9) is Hamiltonian for all $K(z)$ if and only if (2.10) holds. \square

VI.3 First Examples of Symplectic Integrators

Since symplecticity is a characteristic property of Hamiltonian systems (Theorem 2.6), it is natural to search for numerical methods that share this property. Pioneering work on symplectic integration is due to de Vogelaere (1956), Ruth (1983), and Feng Kang (1985). Books on the now well-developed subject are Sanz-Serna & Calvo (1994) and Leimkuhler & Reich (2004).

Definition 3.1. A numerical one-step method is called *symplectic* if the one-step map

$$y_1 = \Phi_h(y_0)$$

is symplectic whenever the method is applied to a smooth Hamiltonian system.



Feng Kang²

² Feng Kang, born: 9 September 1920 in Nanjing (China), died: 17 August 1993 in Beijing; picture obtained from Yuming Shi with the help of Yifa Tang.

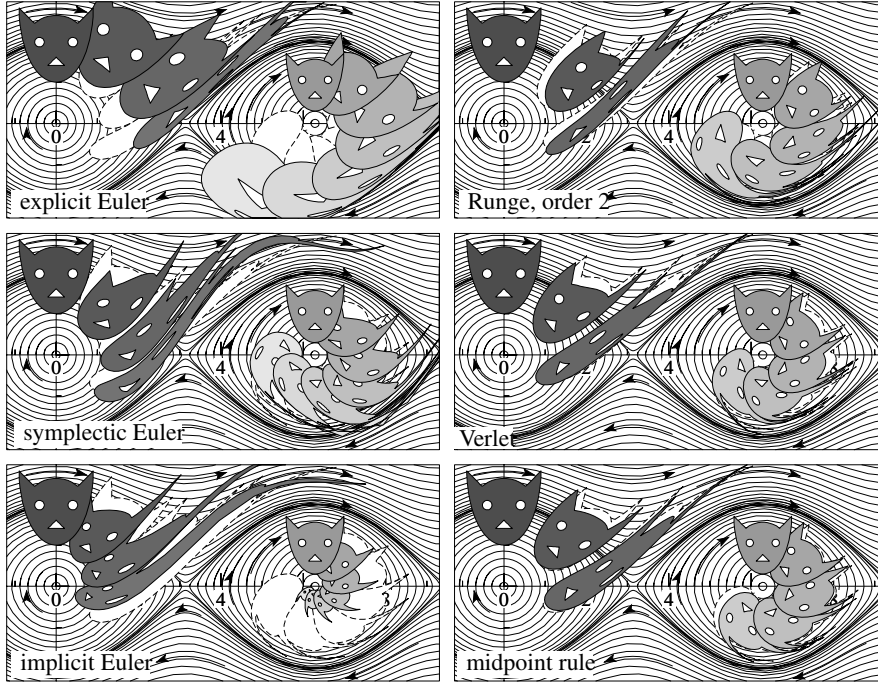


Fig. 3.1. Area preservation of numerical methods for the pendulum; same initial sets as in Fig. 2.2; first order methods (left column): $h = \pi/4$; second order methods (right column): $h = \pi/3$; dashed: exact flow

Example 3.2. We consider the pendulum problem of Example 2.5 with the same initial sets as in Fig. 2.2. We apply six different numerical methods to this problem: the explicit Euler method (I.1.5), the symplectic Euler method (I.1.9), and the implicit Euler method (I.1.6), as well as the second order method of Runge (II.1.3) (the right one), the Störmer–Verlet scheme (I.1.17), and the implicit midpoint rule (I.1.7). For two sets of initial values (p_0, q_0) we compute several steps with step size $h = \pi/4$ for the first order methods, and $h = \pi/3$ for the second order methods. One clearly observes in Fig. 3.1 that the explicit Euler, the implicit Euler and the second order explicit method of Runge are not symplectic (not area preserving). We shall prove below that the other methods are symplectic. A different proof of their symplecticity (using generating functions) will be given in Sect. VI.5.

In the following we show the symplecticity of various numerical methods from Chapters I and II when they are applied to the Hamiltonian system in the variables $y = (p, q)$,

$$\begin{aligned} \dot{p} &= -H_q(p, q) \\ \dot{q} &= H_p(p, q) \end{aligned} \quad \text{or equivalently} \quad \dot{y} = J^{-1} \nabla H(y),$$

where H_p and H_q denote the column vectors of partial derivatives of the Hamiltonian $H(p, q)$ with respect to p and q , respectively.

Theorem 3.3 (de Vogelaere 1956). *The so-called symplectic Euler methods (I.1.9)*

$$\begin{aligned} p_{n+1} &= p_n - hH_q(p_{n+1}, q_n) & \text{or} & & p_{n+1} &= p_n - hH_q(p_n, q_{n+1}) \\ q_{n+1} &= q_n + hH_p(p_{n+1}, q_n) & & & q_{n+1} &= q_n + hH_p(p_n, q_{n+1}) \end{aligned} \quad (3.1)$$

are symplectic methods of order 1.

Proof. We consider only the method to the left of (3.1). Differentiation with respect to (p_n, q_n) yields

$$\begin{pmatrix} I + hH_{qp}^T & 0 \\ -hH_{pp} & I \end{pmatrix} \begin{pmatrix} \frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)} \end{pmatrix} = \begin{pmatrix} I & -hH_{qq} \\ 0 & I + hH_{qp} \end{pmatrix},$$

where the matrices H_{qp}, H_{pp}, \dots of partial derivatives are all evaluated at (p_{n+1}, q_n) . This relation allows us to compute $\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}$ and to check in a straightforward way the symplecticity condition $\left(\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}\right)^T J \left(\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}\right) = J$. \square

The methods (3.1) are implicit for general Hamiltonian systems. For separable $H(p, q) = T(p) + U(q)$, however, both variants turn out to be explicit. It is interesting to mention that there are more general situations where the symplectic Euler methods are explicit. If, for a suitable ordering of the components,

$$\frac{\partial H}{\partial q_i}(p, q) \quad \text{does not depend on } p_j \text{ for } j \geq i, \quad (3.2)$$

then the left method of (3.1) is explicit, and the components of p_{n+1} can be computed one after the other. If, for a possibly different ordering of the components,

$$\frac{\partial H}{\partial p_i}(p, q) \quad \text{does not depend on } q_j \text{ for } j \geq i, \quad (3.3)$$

then the right method of (3.1) is explicit. As an example consider the Hamiltonian

$$H(p_r, p_\varphi, r, \varphi) = \frac{1}{2}(p_r^2 + r^{-2}p_\varphi^2) - r \cos \varphi + (r - 1)^2,$$

which models a spring pendulum in polar coordinates. For the ordering $\varphi < r$, condition (3.2) is fulfilled, and for the inverse ordering $r < \varphi$ condition (3.3). Consequently, both symplectic Euler methods are explicit for this problem. The methods remain explicit if the conditions (3.2) and (3.3) hold for blocks of components instead of single components.

We consider next the extension of the Störmer–Verlet scheme (I.1.17), considered in Table II.2.1.

Theorem 3.4. *The Störmer–Verlet schemes (I.1.17)*

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2}H_q(p_{n+1/2}, q_n) \\ q_{n+1} &= q_n + \frac{h}{2}\left(H_p(p_{n+1/2}, q_n) + H_p(p_{n+1/2}, q_{n+1})\right) \\ p_{n+1} &= p_{n+1/2} - \frac{h}{2}H_q(p_{n+1/2}, q_{n+1}) \end{aligned} \quad (3.4)$$

and

$$\begin{aligned} q_{n+1/2} &= q_n + \frac{h}{2} H_q(p_n, q_{n+1/2}) \\ p_{n+1} &= p_n - \frac{h}{2} \left(H_p(p_n, q_{n+1/2}) + H_p(p_{n+1}, q_{n+1/2}) \right) \\ q_{n+1} &= q_{n+1/2} + \frac{h}{2} H_q(p_{n+1}, q_{n+1/2}) \end{aligned} \quad (3.5)$$

are symplectic methods of order 2.

Proof. This is an immediate consequence of the fact that the Störmer–Verlet scheme is the composition of the two symplectic Euler methods (3.1). Order 2 follows from its symmetry. \square

We note that the Störmer–Verlet methods (3.4) and (3.5) are explicit for separable problems and for Hamiltonians that satisfy both conditions (3.2) and (3.3).

Theorem 3.5. *The implicit midpoint rule*

$$y_{n+1} = y_n + hJ^{-1}\nabla H((y_{n+1} + y_n)/2) \quad (3.6)$$

is a symplectic method of order 2.

Proof. Differentiation of (3.6) yields

$$\left(I - \frac{h}{2} J^{-1} \nabla^2 H \right) \left(\frac{\partial y_{n+1}}{\partial y_n} \right) = \left(I + \frac{h}{2} J^{-1} \nabla^2 H \right).$$

Again it is straightforward to verify that $\left(\frac{\partial y_{n+1}}{\partial y_n} \right)^T J \left(\frac{\partial y_{n+1}}{\partial y_n} \right) = J$. Due to its symmetry, the midpoint rule is known to be of order 2 (see Theorem II.3.2). \square

The next two theorems are a consequence of the fact that the composition of symplectic transformations is again symplectic. They are also used to prove the existence of symplectic methods of arbitrarily high order, and to explain why the theory of composition methods of Chapters II and III is so important for geometric integration.

Theorem 3.6. *Let Φ_h denote the symplectic Euler method (3.1). Then, the composition method (II.4.6) is symplectic for every choice of the parameters α_i, β_i .*

If Φ_h is symplectic and symmetric (e.g., the implicit midpoint rule or the Störmer–Verlet scheme), then the composition method (V.3.8) is symplectic too. \square

Theorem 3.7. *Assume that the Hamiltonian is given by $H(y) = H_1(y) + H_2(y)$, and consider the splitting*

$$\dot{y} = J^{-1}\nabla H(y) = J^{-1}\nabla H_1(y) + J^{-1}\nabla H_2(y).$$

The splitting method (II.5.6) is then symplectic. \square

VI.4 Symplectic Runge–Kutta Methods

The systematic study of symplectic Runge–Kutta methods started around 1988, and a complete characterization has been found independently by Lasagni (1988) (using the approach of generating functions), and by Sanz-Serna (1988) and Suris (1988) (using the ideas of the classical papers of Burrage & Butcher (1979) and Crouzeix (1979) on algebraic stability).

VI.4.1 Criterion of Symplecticity

We follow the approach of Bochev & Scovel (1994), which is based on the following important lemma.

Lemma 4.1. *For Runge–Kutta methods and for partitioned Runge–Kutta methods the following diagram commutes:*

$$\begin{array}{ccc}
 \dot{y} = f(y), \quad y(0) = y_0 & \longrightarrow & \begin{array}{l} \dot{y} = f(y), \quad y(0) = y_0 \\ \dot{\Psi} = f'(y)\Psi, \quad \Psi(0) = I \end{array} \\
 \downarrow \text{method} & & \downarrow \text{method} \\
 \{y_n\} & \longrightarrow & \{y_n, \Psi_n\}
 \end{array}$$

(horizontal arrows mean a differentiation with respect to y_0). Therefore, the numerical result y_n, Ψ_n , obtained from applying the method to the problem augmented by its variational equation, is equal to the numerical solution for $\dot{y} = f(y)$ augmented by its derivative $\Psi_n = \partial y_n / \partial y_0$.

Proof. The result is proved by implicit differentiation. Let us illustrate this for the explicit Euler method

$$y_{n+1} = y_n + hf(y_n).$$

We consider y_n and y_{n+1} as functions of y_0 , and we differentiate with respect to y_0 the equation defining the numerical method. For the Euler method this gives

$$\frac{\partial y_{n+1}}{\partial y_0} = \frac{\partial y_n}{\partial y_0} + hf'(y_n) \frac{\partial y_n}{\partial y_0},$$

which is exactly the relation that we get from applying the method to the variational equation. Since $\partial y_0 / \partial y_0 = I$, we have $\partial y_n / \partial y_0 = \Psi_n$ for all n . \square

The main observation now is that the symplecticity condition (2.6) is a quadratic first integral of the variational equation: we write the Hamiltonian system together with its variational equation as

$$\dot{y} = J^{-1} \nabla H(y), \quad \dot{\Psi} = J^{-1} \nabla^2 H(y) \Psi. \quad (4.1)$$

It follows from

$$(J^{-1}\nabla^2 H(y)\Psi)^T J\Psi + \Psi^T J(J^{-1}\nabla^2 H(y)\Psi) = 0$$

(see also the proof of Theorem 2.4) that $\Psi^T J\Psi$ is a quadratic first integral of the augmented system (4.1).

Therefore, every Runge–Kutta method that preserves quadratic first integrals, is a symplectic method. From Theorem IV.2.1 and Theorem IV.2.2 we thus obtain the following results.

Theorem 4.2. *The Gauss collocation methods of Sect. II.1.3 are symplectic.* \square

Theorem 4.3. *If the coefficients of a Runge–Kutta method satisfy*

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \quad \text{for all } i, j = 1, \dots, s, \quad (4.2)$$

then it is symplectic. \square

Similar to the situation in Theorem V.2.4, diagonally implicit, symplectic Runge–Kutta methods are composition methods.

Theorem 4.4. *A diagonally implicit Runge–Kutta method satisfying the symplecticity condition (4.2) and $b_i \neq 0$ is equivalent to the composition*

$$\Phi_{b_s h}^M \circ \dots \circ \Phi_{b_2 h}^M \circ \Phi_{b_1 h}^M,$$

where Φ_h^M stands for the implicit midpoint rule.

Proof. For $i = j$ condition (4.2) gives $a_{ii} = b_i/2$ and, together with $a_{ji} = 0$ (for $i > j$), implies $a_{ij} = b_j$. This proves the statement. \square

The assumption “ $b_i \neq 0$ ” is not restrictive in the sense that for diagonally implicit Runge–Kutta methods satisfying (4.2) the internal stages corresponding to “ $b_i = 0$ ” do not influence the numerical result and can be removed.

To understand the symplecticity of partitioned Runge–Kutta methods, we write the solution Ψ of the variational equation as

$$\Psi = \begin{pmatrix} \Psi^p \\ \Psi^q \end{pmatrix}.$$

Then, the Hamiltonian system together with its variational equation (4.1) is a partitioned system with variables (p, Ψ^p) and (q, Ψ^q) . Every component of

$$\Psi^T J\Psi = (\Psi^p)^T \Psi^q - (\Psi^q)^T \Psi^p$$

is of the form (IV.2.5), so that Theorem IV.2.3 and Theorem IV.2.4 yield the following results.

Theorem 4.5. *The Lobatto IIIA - IIIB pair is a symplectic method.* \square

Theorem 4.6. *If the coefficients of a partitioned Runge–Kutta method (II.2.2) satisfy*

$$b_i \hat{a}_{ij} + \hat{b}_j a_{ji} = b_i \hat{b}_j \quad \text{for } i, j = 1, \dots, s, \quad (4.3)$$

$$b_i = \hat{b}_i \quad \text{for } i = 1, \dots, s, \quad (4.4)$$

then it is symplectic.

If the Hamiltonian is of the form $H(p, q) = T(p) + U(q)$, i.e., it is separable, then the condition (4.3) alone implies the symplecticity of the numerical flow. \square

We have seen in Sect. V.2.2 that within the class of partitioned Runge–Kutta methods it is possible to get explicit, symmetric methods for separable systems $\dot{y} = f(z)$, $\dot{z} = g(y)$. A similar result holds for symplectic methods. However, as in Theorem V.2.6, such methods are not more general than composition or splitting methods as considered in Sect. II.5. This has first been observed by Okunbor & Skeel (1992).

Theorem 4.7. *Consider a partitioned Runge–Kutta method based on two diagonally implicit methods (i.e., $a_{ji} = \hat{a}_{ji} = 0$ for $i > j$), assume $a_{ii} \cdot \hat{a}_{ii} = 0$ for all i , and apply it to a separable Hamiltonian system with $H(p, q) = T(p) + U(q)$. If (4.3) holds, then the numerical result is the same as that obtained from the splitting method (II.5.6).*

By (II.5.8), such a method is equivalent to a composition of symplectic Euler steps.

Proof. We first notice that the stage values $k_i = f(Z_i)$ (for i with $b_i = 0$) and $\ell_i = g(Y_i)$ (for i with $\hat{b}_i = 0$) do not influence the numerical solution and can be removed. This yields a scheme with non-zero b_i and \hat{b}_i , but with possibly non-square matrices (a_{ij}) and (\hat{a}_{ij}) .

Since the method is explicit for separable problems, one of the reduced matrices (a_{ij}) or (\hat{a}_{ij}) has a row consisting only of zeros. Assume that it is the first row of (a_{ij}) , so that $a_{1j} = 0$ for all j . The symplecticity condition thus implies $\hat{a}_{i1} = \hat{b}_1 \neq 0$ for all $i \geq 1$, and $a_{i1} = b_1 \neq 0$ for $i \geq 2$. This then yields $\hat{a}_{22} \neq 0$, because otherwise the first two stages of (\hat{a}_{ij}) would be identical and one could be removed. By our assumption we get $a_{22} = 0$, $\hat{a}_{i2} = \hat{b}_2 \neq 0$ for $i \geq 2$, and $a_{i2} = b_2$ for $i \geq 3$. Continuing this procedure we see that the method becomes

$$\dots \circ \varphi_{\hat{b}_2 h}^{[2]} \circ \varphi_{b_2 h}^{[1]} \circ \varphi_{\hat{b}_1 h}^{[2]} \circ \varphi_{b_1 h}^{[1]},$$

where $\varphi_t^{[1]}$ and $\varphi_t^{[2]}$ are the exact flows corresponding to the Hamiltonians $T(p)$ and $U(q)$, respectively. \square

The necessity of the conditions of Theorem 4.3 and Theorem 4.6 for symplectic (partitioned) Runge–Kutta methods will be discussed at the end of this chapter in Sect. VI.7.4.

A second order differential equation $\ddot{y} = g(y)$, augmented by its variational equation, is again of this special form. Furthermore, the diagram of Lemma 4.1 commutes for Nyström methods, so that Theorem IV.2.5 yields the following result originally obtained by Suris (1988, 1989).

Theorem 4.8. *If the coefficients of a Nyström method (IV.2.11) satisfy*

$$\begin{aligned} \beta_i &= b_i(1 - c_i) & \text{for } i = 1, \dots, s, \\ b_i(\beta_j - a_{ij}) &= b_j(\beta_i - a_{ji}) & \text{for } i, j = 1, \dots, s, \end{aligned} \quad (4.5)$$

then it is symplectic. □

VI.4.2 Connection Between Symplectic and Symmetric Methods

There exist symmetric methods that are not symplectic, and there exist symplectic methods that are not symmetric. For example, the *trapezoidal rule*

$$y_1 = y_0 + \frac{h}{2} (f(y_0) + f(y_1)) \quad (4.6)$$

is symmetric, but it does not satisfy the condition (4.2) for symplecticity. In fact, this is true of all Lobatto IIIA methods (see Example II.2.2). On the other hand, any composition $\Phi_{\gamma_1 h} \circ \Phi_{\gamma_2 h}$ ($\gamma_1 + \gamma_2 = 1$) of symplectic methods is symplectic but symmetric only if $\gamma_1 = \gamma_2$.

However, for (non-partitioned) Runge–Kutta methods and for quadratic Hamiltonians $H(y) = \frac{1}{2} y^T C y$ (C is a symmetric real matrix), where the corresponding system (2.5) is linear,

$$\dot{y} = J^{-1} C y, \quad (4.7)$$

we shall see that both concepts are equivalent.

A Runge–Kutta method, applied with step size h to a linear system $\dot{y} = Ly$, is equivalent to

$$y_1 = R(hL)y_0, \quad (4.8)$$

where the rational function $R(z)$ is given by

$$R(z) = 1 + z b^T (I - zA)^{-1} \mathbb{1}, \quad (4.9)$$

$A = (a_{ij})$, $b^T = (b_1, \dots, b_s)$, and $\mathbb{1}^T = (1, \dots, 1)$. The function $R(z)$ is called the *stability function* of the method, and it is familiar to us from the study of stiff differential equations (see e.g., Hairer & Wanner (1996), Chap. IV.3).

For the explicit Euler method, the implicit Euler method and the implicit mid-point rule, the stability function $R(z)$ is given by

$$1 + z, \quad \frac{1}{1 - z}, \quad \frac{1 + z/2}{1 - z/2}.$$

Theorem 4.9. *For Runge–Kutta methods the following statements are equivalent:*

- *the method is symmetric for linear problems $\dot{y} = Ly$;*
- *the method is symplectic for problems (4.7) with symmetric C ;*
- *the stability function satisfies $R(-z)R(z) = 1$ for all complex z .*

Proof. The method $y_1 = R(hL)y_0$ is symmetric, if and only if $y_0 = R(-hL)y_1$ holds for all initial values y_0 . But this is equivalent to $R(-hL)R(hL) = I$.

Since $\Phi'_h(y_0) = R(hL)$, symplecticity of the method for the problem (4.7) is defined by $R(hJ^{-1}C)^T J R(hJ^{-1}C) = J$. For $R(z) = P(z)/Q(z)$ this is equivalent to

$$P(hJ^{-1}C)^T J P(hJ^{-1}C) = Q(hJ^{-1}C)^T J Q(hJ^{-1}C). \quad (4.10)$$

By the symmetry of C , the matrix $L := J^{-1}C$ satisfies $L^T J = -JL$ and hence also $(L^k)^T J = J(-L)^k$ for $k = 0, 1, 2, \dots$. Consequently, (4.10) is equivalent to

$$P(-hJ^{-1}C)P(hJ^{-1}C) = Q(-hJ^{-1}C)Q(hJ^{-1}C),$$

which is nothing other than $R(-hJ^{-1}C)R(hJ^{-1}C) = I$. \square

VI.5 Generating Functions

... by which the study of the motions of all free systems of attracting or repelling points is reduced to the search and differentiation of one central relation, or characteristic function. (W.R. Hamilton 1834)

Professor Hamilton hat ... das merkwürdige Resultat gefunden, dass ... sich die Integralgleichungen der Bewegung ... sämtlich durch die partiellen Differentialquotienten einer einzigen Function darstellen lassen. (C.G.J. Jacobi 1837)

We enter here the second heaven of Hamiltonian theory, the realm of partial differential equations and generating functions. The starting point of this theory was the discovery of Hamilton that the motion of the system is completely described by a “characteristic” function S , and that S is the solution of a partial differential equation, now called the *Hamilton–Jacobi differential equation*.

It was noticed later, especially by Siegel (see Siegel & Moser 1971, §3), that such a function S is directly connected to any symplectic map. It received the name *generating function*.

VI.5.1 Existence of Generating Functions

We now consider a fixed Hamiltonian system and a fixed time interval and denote by the column vectors p and q the *initial values* p_1, \dots, p_d and q_1, \dots, q_d at t_0 of a trajectory. The *final values* at t_1 are written as P and Q . We thus have a mapping $(p, q) \mapsto (P, Q)$ which, as we know, is symplectic on an open set U .

The following results are conveniently formulated in the notation of differential forms. For a function F we denote by $dF = F'$ its (Fréchet) derivative. We denote by $dq = (dq_1, \dots, dq_d)^T$ the derivative of the coordinate projection $(p, q) \mapsto q$.

Theorem 5.1. *A mapping $\varphi : (p, q) \mapsto (P, Q)$ is symplectic if and only if there exists locally a function $S(p, q)$ such that*

$$P^T dQ - p^T dq = dS. \quad (5.1)$$

This means that $P^T dQ - p^T dq$ is a total differential.

Proof. We split the Jacobian of φ into the natural 2×2 block matrix

$$\frac{\partial(P, Q)}{\partial(p, q)} = \begin{pmatrix} P_p & P_q \\ Q_p & Q_q \end{pmatrix}.$$

Inserting this into (2.6) and multiplying out shows that the three conditions

$$P_p^T Q_p = Q_p^T P_p, \quad P_p^T Q_q - I = Q_p^T P_q, \quad Q_q^T P_q = P_q^T Q_q \quad (5.2)$$

are equivalent to symplecticity. We now insert $dQ = Q_p dp + Q_q dq$ into the left-hand side of (5.1) and obtain

$$\left(P^T Q_p, P^T Q_q - p^T \right) \begin{pmatrix} dp \\ dq \end{pmatrix} = \begin{pmatrix} Q_p^T P & Q_q^T P - p \end{pmatrix}^T \begin{pmatrix} dp \\ dq \end{pmatrix}.$$

To apply the Integrability Lemma 2.7, we just have to verify the symmetry of the Jacobian of the coefficient vector,

$$\begin{pmatrix} Q_p^T P_p & Q_p^T P_q \\ Q_q^T P_p - I & Q_q^T P_q \end{pmatrix} + \sum_i P_i \frac{\partial^2 Q_i}{\partial(p, q)^2}. \quad (5.3)$$

Since the Hessians of Q_i are symmetric anyway, it is immediately clear that the symmetry of the matrix (5.3) is equivalent to the symplecticity conditions (5.2). \square

Reconstruction of the Symplectic Map from S . Up to now we have considered all functions as depending on p and q . The essential idea now is to introduce new coordinates; namely (5.1) suggests using $z = (q, Q)$ instead of $y = (p, q)$. This is a well-defined local change of coordinates $y = \psi(z)$ if p can be expressed in terms of the coordinates (q, Q) , which is possible by the implicit function theorem if $\frac{\partial Q}{\partial p}$ is invertible. Abusing our notation we again write $S(q, Q)$ for the transformed function $S(\psi(z))$. Then, by comparing the coefficients of $dS = \frac{\partial S(q, Q)}{\partial q} dq + \frac{\partial S(q, Q)}{\partial Q} dQ$ with (5.1), we arrive at³

$$P = \frac{\partial S}{\partial Q}(q, Q), \quad p = -\frac{\partial S}{\partial q}(q, Q). \quad (5.4)$$

If the transformation $(p, q) \mapsto (P, Q)$ is symplectic, then it can be reconstructed from the scalar function $S(q, Q)$ by the relations (5.4). By Theorem 5.1 the converse

³ On the right-hand side we should have put the gradient $\nabla_Q S = (\partial S / \partial Q)^T$. We shall not make this distinction between row and column vectors when there is no danger of confusion.

is also true: any sufficiently smooth and nondegenerate function $S(q, Q)$ “generates” via (5.4) a symplectic mapping $(p, q) \mapsto (P, Q)$. This gives us a powerful tool for creating symplectic methods.

Mixed-Variable Generating Functions. Another often useful choice of coordinates for generating symplectic maps are the mixed variables (P, q) . For any continuously differentiable function $\hat{S}(P, q)$ we clearly have $d\hat{S} = \frac{\partial \hat{S}}{\partial P} dP + \frac{\partial \hat{S}}{\partial q} dq$. On the other hand, since $d(P^T Q) = P^T dQ + Q^T dP$, the symplecticity condition (5.1) can be rewritten as $Q^T dP + p^T dq = d(Q^T P - S)$ for some function S . It therefore follows from Theorem 5.1 that the equations

$$Q = \frac{\partial \hat{S}}{\partial P}(P, q), \quad p = \frac{\partial \hat{S}}{\partial q}(P, q) \quad (5.5)$$

define (locally) a symplectic map $(p, q) \mapsto (P, Q)$ if $\partial^2 \hat{S} / \partial P \partial q$ is invertible.

Example 5.2. Let $Q = \chi(q)$ be a change of position coordinates. With the generating function $\hat{S}(P, q) = P^T \chi(q)$ we obtain via (5.5) an extension to a symplectic mapping $(p, q) \mapsto (P, Q)$. The conjugate variables are thus related by $p = \chi'(q)^T P$.

Mappings Close to the Identity. We are mainly interested in the situation where the mapping $(p, q) \mapsto (P, Q)$ is close to the identity. In this case, the choices (p, Q) or (P, q) or $((P + p)/2, (Q + q)/2)$ of independent variables are convenient and lead to the following characterizations.

Lemma 5.3. Let $(p, q) \mapsto (P, Q)$ be a smooth transformation, close to the identity. It is symplectic if and only if one of the following conditions holds locally:

- $Q^T dP + p^T dq = d(P^T q + S^1)$ for some function $S^1(P, q)$;
- $P^T dQ + q^T dp = d(p^T Q - S^2)$ for some function $S^2(p, Q)$;
- $(Q - q)^T d(P + p) - (P - p)^T d(Q + q) = 2 dS^3$
for some function $S^3((P + p)/2, (Q + q)/2)$.

Proof. The first characterization follows from the discussion before formula (5.5) if we put S^1 such that $P^T q + S^1 = \hat{S} = Q^T P - S$. For the second characterization we use $d(p^T q) = p^T dq + q^T dp$ and the same arguments as before. The last one follows from the fact that (5.1) is equivalent to $(Q - q)^T d(P + p) - (P - p)^T d(Q + q) = d((P + p)^T (Q - q) - 2S)$. \square

The generating functions S^1 , S^2 , and S^3 have been chosen such that we obtain the identity mapping when they are replaced with zero. Comparing the coefficient functions of dq and dP in the first characterization of Lemma 5.3, we obtain

$$p = P + \frac{\partial S^1}{\partial q}(P, q), \quad Q = q + \frac{\partial S^1}{\partial P}(P, q). \quad (5.6)$$

Whatever the scalar function $S^1(P, q)$ is, the relation (5.6) defines a symplectic transformation $(p, q) \mapsto (P, Q)$. For $S^1(P, q) := hH(P, q)$ we recognize the symplectic Euler method (I.1.9). This is an elegant proof of the symplecticity of this method. The second characterization leads to the adjoint of the symplectic Euler method.

The third characterization of Lemma 5.3 can be written as

$$\begin{aligned} P &= p - \partial_2 S^3((P+p)/2, (Q+q)/2), \\ Q &= q + \partial_1 S^3((P+p)/2, (Q+q)/2), \end{aligned} \quad (5.7)$$

which, for $S^3 = hH$, is nothing other than the implicit midpoint rule (I.1.7) applied to a Hamiltonian system. We have used the notation ∂_1 and ∂_2 for the derivative with respect to the first and second argument, respectively. The system (5.7) can also be written in compact form as

$$Y = y + J^{-1} \nabla S^3((Y+y)/2), \quad (5.8)$$

where $Y = (P, Q)$, $y = (p, q)$, $S^3(w) = S^3(u, v)$ with $w = (u, v)$, and J is the matrix of (2.3).

VI.5.2 Generating Function for Symplectic Runge–Kutta Methods

We have just seen that all symplectic transformations can be written in terms of generating functions. What are these generating functions for symplectic Runge–Kutta methods? The following result, proved by Lasagni in an unpublished manuscript (with the same title as the note Lasagni (1988)), gives an alternative proof for Theorem 4.3.

Theorem 5.4. *Suppose that*

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \quad \text{for all } i, j \quad (5.9)$$

(see Theorem 4.3). Then, the Runge–Kutta method

$$\begin{aligned} P &= p - h \sum_{i=1}^s b_i H_q(P_i, Q_i), & P_i &= p - h \sum_{j=1}^s a_{ij} H_q(P_j, Q_j), \\ Q &= q + h \sum_{i=1}^s b_i H_p(P_i, Q_i), & Q_i &= q + h \sum_{j=1}^s a_{ij} H_p(P_j, Q_j) \end{aligned} \quad (5.10)$$

can be written as (5.6) with

$$S^1(P, q, h) = h \sum_{i=1}^s b_i H(P_i, Q_i) - h^2 \sum_{i,j=1}^s b_i a_{ij} H_q(P_i, Q_i)^T H_p(P_j, Q_j). \quad (5.11)$$

Proof. We first differentiate $S^1(P, q, h)$ with respect to q . Using the abbreviations $H[i] = H(P_i, Q_i)$, $H_p[i] = H_p(P_i, Q_i)$, \dots , we obtain

$$\begin{aligned} \frac{\partial}{\partial q} \left(\sum_i b_i H[i] \right) &= \sum_i b_i H_p[i]^T \left(\frac{\partial p}{\partial q} - h \sum_j a_{ij} \frac{\partial}{\partial q} H_q[j] \right) \\ &+ \sum_i b_i H_q[i]^T \left(I + h \sum_j a_{ij} \frac{\partial}{\partial q} H_p[j] \right). \end{aligned}$$

With

$$0 = \frac{\partial p}{\partial q} - h \sum_j b_j \frac{\partial}{\partial q} H_q[j]$$

(this is obtained by differentiating the first relation of (5.10)), Leibniz' rule

$$\frac{\partial}{\partial q} (H_q[i]^T H_p[j]) = H_q[i]^T \frac{\partial}{\partial q} H_p[j] + H_p[j]^T \frac{\partial}{\partial q} H_q[i]$$

and the condition (5.9) therefore yield the first relation of

$$\frac{\partial S^1(P, q, h)}{\partial q} = h \sum_i b_i H_q[i], \quad \frac{\partial S^1(P, q, h)}{\partial P} = h \sum_i b_i H_p[i].$$

The second relation is proved in the same way. This shows that the Runge–Kutta formulas (5.10) are equivalent to (5.6). \square

It is interesting to note that, whereas Lemma 5.3 guarantees the *local* existence of a generating function S^1 , the explicit formula (5.11) shows that for Runge–Kutta methods this generating function is *globally* defined. This means that it is well-defined in the same region where the Hamiltonian $H(p, q)$ is defined.

Theorem 5.5. *A partitioned Runge–Kutta method (II.2.2), satisfying the symplecticity conditions (4.3) and (4.4), is equivalent to (5.6) with*

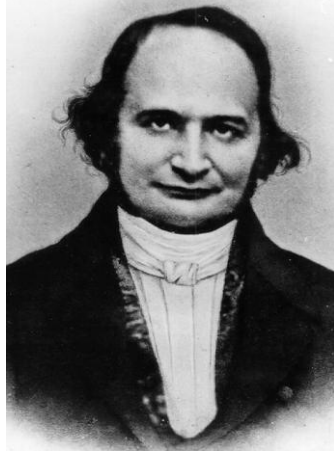
$$S^1(P, q, h) = h \sum_{i=1}^s b_i H(P_i, Q_i) - h^2 \sum_{i,j=1}^s b_i \hat{a}_{ij} H_q(P_i, Q_i)^T H_p(P_j, Q_j).$$

If the Hamiltonian is of the form $H(p, q) = T(p) + U(q)$, i.e., it is separable, then the condition (4.3) alone implies that the method is of the form (5.6) with

$$S^1(P, q, h) = h \sum_{i=1}^s \left(b_i U(Q_i) + \hat{b}_i T(P_i) \right) - h^2 \sum_{i,j=1}^s b_i \hat{a}_{ij} U_q(Q_i)^T T_p(P_j).$$

Proof. This is a straightforward extension of the proof of the previous theorem. \square

VI.5.3 The Hamilton–Jacobi Partial Differential Equation

C.G.J. Jacobi⁴

We now return to the above construction of S for a symplectic transformation $(p, q) \mapsto (P, Q)$ (see Theorem 5.1). This time, however, we imagine the point $P(t), Q(t)$ to move in the flow of the Hamiltonian system (1.7). We wish to determine a smooth generating function $S(q, Q, t)$, now also depending on t , which generates via (5.4) the symplectic map $(p, q) \mapsto (P(t), Q(t))$ of the *exact flow* of the Hamiltonian system.

In accordance with equation (5.4) we have to satisfy

$$\begin{aligned} P_i(t) &= \frac{\partial S}{\partial Q_i}(q, Q(t), t), \\ p_i &= -\frac{\partial S}{\partial q_i}(q, Q(t), t). \end{aligned} \quad (5.12)$$

Differentiating the second relation with respect to t yields

$$0 = \frac{\partial^2 S}{\partial q_i \partial t}(q, Q(t), t) + \sum_{j=1}^d \frac{\partial^2 S}{\partial q_i \partial Q_j}(q, Q(t), t) \cdot \dot{Q}_j(t) \quad (5.13)$$

$$= \frac{\partial^2 S}{\partial q_i \partial t}(q, Q(t), t) + \sum_{j=1}^d \frac{\partial^2 S}{\partial q_i \partial Q_j}(q, Q(t), t) \cdot \frac{\partial H}{\partial P_j}(P(t), Q(t)) \quad (5.14)$$

where we have inserted the second equation of (1.7) for \dot{Q}_j . Then, using the chain rule, this equation simplifies to

$$\frac{\partial}{\partial q_i} \left(\frac{\partial S}{\partial t} + H \left(\frac{\partial S}{\partial Q_1}, \dots, \frac{\partial S}{\partial Q_d}, Q_1, \dots, Q_d \right) \right) = 0. \quad (5.15)$$

This motivates the following surprisingly simple relation.

Theorem 5.6. *If $S(q, Q, t)$ is a smooth solution of the partial differential equation*

$$\frac{\partial S}{\partial t} + H \left(\frac{\partial S}{\partial Q_1}, \dots, \frac{\partial S}{\partial Q_d}, Q_1, \dots, Q_d \right) = 0 \quad (5.16)$$

with initial values satisfying $\frac{\partial S}{\partial q_i}(q, q, 0) + \frac{\partial S}{\partial Q_i}(q, q, 0) = 0$, and if the matrix $\left(\frac{\partial^2 S}{\partial q_i \partial Q_j} \right)$ is invertible, then the map $(p, q) \mapsto (P(t), Q(t))$ defined by (5.12) is the flow $\varphi_t(p, q)$ of the Hamiltonian system (1.7).

Equation (5.16) is called the “Hamilton–Jacobi partial differential equation”.

⁴ Carl Gustav Jacob Jacobi, born: 10 December 1804 in Potsdam (near Berlin), died: 18 February 1851 in Berlin.

Proof. The invertibility of the matrix $(\frac{\partial^2 S}{\partial q_i \partial Q_j})$ and the implicit function theorem imply that the mapping $(p, q) \mapsto (P(t), Q(t))$ is well-defined by (5.12), and, by differentiation, that (5.13) is true as well.

Since, by hypothesis, $S(q, Q, t)$ is a solution of (5.16), the equations (5.15) and hence also (5.14) are satisfied. Subtracting (5.13) and (5.14), and once again using the invertibility of the matrix $(\frac{\partial^2 S}{\partial q_i \partial Q_j})$, we see that necessarily $\dot{Q}(t) = H_p(P(t), Q(t))$. This proves the validity of the second equation of the Hamiltonian system (1.7).

The first equation of (1.7) is obtained as follows: differentiate the first relation of (5.12) with respect to t and the Hamilton–Jacobi equation (5.16) with respect to Q_i , then eliminate the term $\frac{\partial^2 S}{\partial Q_i \partial t}$. Using $\dot{Q}(t) = H_p(P(t), Q(t))$, this leads in a straightforward way to $\dot{P}(t) = -H_q(P(t), Q(t))$. The condition on the initial values of S ensures that $(P(0), Q(0)) = (p, q)$. \square

In the hands of Jacobi (1842), this equation turned into a powerful tool for the analytic integration of many difficult problems. One has, in fact, to find a solution of (5.16) which contains sufficiently many parameters. This is often possible with the method of separation of variables. An example is presented in Exercise 11.

Hamilton–Jacobi Equation for S^1 , S^2 , and S^3 . We now express the Hamilton–Jacobi differential equation in the coordinates used in Lemma 5.3. In these coordinates it is also possible to prescribe initial values for S at $t = 0$.

From the proof of Lemma 5.3 we know that the generating functions in the variables (q, Q) and (P, q) are related by

$$S^1(P, q, t) = P^T(Q - q) - S(q, Q, t). \quad (5.17)$$

We consider P, q, t as independent variables, and we differentiate this relation with respect to t . Using the first relation of (5.12) this gives

$$\frac{\partial S^1}{\partial t}(P, q, t) = P^T \frac{\partial Q}{\partial t} - \frac{\partial S}{\partial Q}(q, Q, t) \frac{\partial Q}{\partial t} - \frac{\partial S}{\partial t}(q, Q, t) = -\frac{\partial S}{\partial t}(q, Q, t).$$

Differentiating (5.17) with respect to P yields

$$\frac{\partial S^1}{\partial P}(P, q, t) = Q - q + P^T \frac{\partial Q}{\partial P} - \frac{\partial S}{\partial Q}(q, Q, t) \frac{\partial Q}{\partial P} = Q - q.$$

Inserting $\frac{\partial S}{\partial Q} = P$ and $Q = q + \frac{\partial S^1}{\partial P}$ into the Hamilton–Jacobi equation (5.16) we are led to the equation of the following theorem.

Theorem 5.7. *If $S^1(P, q, t)$ is a solution of the partial differential equation*

$$\frac{\partial S^1}{\partial t}(P, q, t) = H\left(P, q + \frac{\partial S^1}{\partial P}(P, q, t)\right), \quad S^1(P, q, 0) = 0, \quad (5.18)$$

then the mapping $(p, q) \mapsto (P(t), Q(t))$, defined by (5.6), is the exact flow of the Hamiltonian system (1.7).

Proof. Whenever the mapping $(p, q) \mapsto (P(t), Q(t))$ can be written as (5.12) with a function $S(q, Q, t)$, and when the invertibility assumption of Theorem 5.6 holds, the proof is done by the above calculations. Since our mapping, for $t = 0$, reduces to the identity and cannot be written as (5.12), we give a direct proof.

Let $S^1(P, q, t)$ be given by the Hamilton–Jacobi equation (5.18), and assume that $(p, q) \mapsto (P, Q) = (P(t), Q(t))$ is the transformation given by (5.6). Differentiation of the first relation of (5.6) with respect to time t and using (5.18) yields⁵

$$\left(I + \frac{\partial^2 S^1}{\partial P \partial q}(P, q, t)\right) \dot{P} = -\frac{\partial^2 S^1}{\partial t \partial q}(P, q, t) = -\left(I + \frac{\partial^2 S^1}{\partial P \partial q}(P, q, t)\right) \frac{\partial H}{\partial Q}(P, Q).$$

Differentiation of the second relation of (5.6) gives

$$\begin{aligned} \dot{Q} &= \frac{\partial^2 S^1}{\partial t \partial P}(P, q, t) + \frac{\partial^2 S^1}{\partial P^2}(P, q, t) \dot{P} \\ &= \frac{\partial H}{\partial P}(P, Q) + \frac{\partial^2 S^1}{\partial P^2}(P, q, t) \left(\frac{\partial H}{\partial Q}(P, Q) + \dot{P} \right). \end{aligned}$$

Consequently, $\dot{P} = -\frac{\partial H}{\partial Q}(P, Q)$ and $\dot{Q} = \frac{\partial H}{\partial P}(P, Q)$, so that $(P(t), Q(t)) = \varphi_t(p, q)$ is the exact flow of the Hamiltonian system. \square

Writing the Hamilton–Jacobi differential equation in the variables $(P + p)/2$, $(Q + q)/2$ gives the following formula.

Theorem 5.8. *Assume that $S^3(u, v, t)$ is a solution of*

$$\frac{\partial S^3}{\partial t}(u, v, t) = H\left(u - \frac{1}{2} \frac{\partial S^3}{\partial v}(u, v, t), v + \frac{1}{2} \frac{\partial S^3}{\partial u}(u, v, t)\right) \quad (5.19)$$

with initial condition $S^3(u, v, 0) = 0$. Then, the exact flow $\varphi_t(p, q)$ of the Hamiltonian system (1.7) satisfies the system (5.7).

Proof. As in the proof of Theorem 5.7, one considers the transformation $(p, q) \mapsto (P(t), Q(t))$ defined by (5.7), and then checks by differentiation that $(P(t), Q(t))$ is a solution of the Hamiltonian system (1.7). \square

Writing $w = (u, v)$ and using the matrix J of (2.3), the Hamilton–Jacobi equation (5.19) can also be written as

$$\frac{\partial S^3}{\partial t}(w, t) = H\left(w + \frac{1}{2} J^{-1} \nabla S^3(w, t)\right), \quad S^3(w, 0) = 0. \quad (5.20)$$

The solution of (5.20) is anti-symmetric in t , i.e.,

$$S^3(w, -t) = -S^3(w, t). \quad (5.21)$$

⁵ Due to an inconsistent notation of the partial derivatives $\frac{\partial H}{\partial Q}$, $\frac{\partial S^1}{\partial q}$ as column or row vectors, this formula may be difficult to read. Use indices instead of matrices in order to check its correctness.

This can be seen as follows: let $\varphi_t(w)$ be the exact flow of the Hamiltonian system $\dot{y} = J^{-1}\nabla H(y)$. Because of (5.8), $S^3(w, t)$ is defined by

$$\varphi_t(w) - w = J^{-1}\nabla S^3((\varphi_t(w) + w)/2, t).$$

Replacing t with $-t$ and then w with $\varphi_t(w)$ we get from $\varphi_{-t}(\varphi_t(t)) = w$ that

$$w - \varphi_t(w) = J^{-1}\nabla S^3((w + \varphi_t(w))/2, -t).$$

Hence $S^3(w, t)$ and $-S^3(w, -t)$ are generating functions of the same symplectic transformation. Since generating functions are unique up to an additive constant (because $dS = 0$ implies $S = \text{Const}$), the anti-symmetry (5.21) follows from the initial condition $S^3(w, 0) = 0$.

VI.5.4 Methods Based on Generating Functions

To construct symplectic numerical methods of high order, Feng Kang (1986), Feng Kang, Wu, Qin & Wang (1989) and Channell & Scovel (1990) proposed computing an approximate solution of the Hamilton–Jacobi equation. For this one inserts the ansatz

$$S^1(P, q, t) = tG_1(P, q) + t^2G_2(P, q) + t^3G_3(P, q) + \dots$$

into (5.18), and compares like powers of t . This yields

$$\begin{aligned} G_1(P, q) &= H(P, q), \\ G_2(P, q) &= \frac{1}{2} \left(\frac{\partial H}{\partial P} \frac{\partial H}{\partial q} \right) (P, q), \\ G_3(P, q) &= \frac{1}{6} \left(\frac{\partial^2 H}{\partial P^2} \left(\frac{\partial H}{\partial q} \right)^2 + \frac{\partial^2 H}{\partial P \partial q} \frac{\partial H}{\partial P} \frac{\partial H}{\partial q} + \frac{\partial^2 H}{\partial q^2} \left(\frac{\partial H}{\partial P} \right)^2 \right) (P, q). \end{aligned}$$

If we use the truncated series

$$S^1(P, q) = hG_1(P, q) + h^2G_2(P, q) + \dots + h^rG_r(P, q) \quad (5.22)$$

and insert it into (5.6), the transformation $(p, q) \mapsto (P, Q)$ defines a symplectic one-step method of order r . Symplecticity follows at once from Lemma 5.3 and order r is a consequence of the fact that the truncation of $S^1(P, q)$ introduces a perturbation of size $\mathcal{O}(h^{r+1})$ in (5.18). We remark that for $r \geq 2$ the methods obtained require the computation of higher derivatives of $H(p, q)$, and for separable Hamiltonians $H(p, q) = T(p) + U(q)$ they are no longer explicit (compared to the symplectic Euler method (3.1)).

The same approach applied to the third characterization of Lemma 5.3 yields

$$S^3(w, h) = hG_1(w) + h^3G_3(w) + \dots + h^{2r-1}G_{2r-1}(w),$$

where $G_1(w) = H(w)$,

$$G_3(w) = \frac{1}{24} \nabla^2 H(w) \left(J^{-1} \nabla H(w), J^{-1} \nabla H(w) \right),$$

and further $G_j(w)$ can be obtained by comparing like powers of h in (5.20). In this way we get symplectic methods of order $2r$. Since $S^3(w, h)$ has an expansion in odd powers of h , the resulting method is symmetric.

The Approach of Miesbach & Pesch. With the aim of avoiding higher derivatives of the Hamiltonian in the numerical method, Miesbach & Pesch (1992) propose considering generating functions of the form

$$S^3(w, h) = h \sum_{i=1}^s b_i H \left(w + h c_i J^{-1} \nabla H(w) \right), \quad (5.23)$$

and to determine the free parameters b_i, c_i in such a way that the function of (5.23) agrees with the solution of the Hamilton–Jacobi equation (5.20) up to a certain order. For $b_{s+1-i} = b_i$ and $c_{s+1-i} = -c_i$ this function satisfies $S^3(w, -h) = -S^3(w, h)$, so that the resulting method is symmetric. A straightforward computation shows that it yields a method of order 4 if

$$\sum_{i=1}^s b_i = 1, \quad \sum_{i=1}^s b_i c_i^2 = \frac{1}{12}.$$

For $s = 3$, these equations are fulfilled for $b_1 = b_3 = 5/18$, $b_2 = 4/9$, $c_1 = -c_3 = \sqrt{15}/10$, and $c_2 = 0$. Since the function S^3 of (5.23) has to be inserted into (5.20), these methods still need second derivatives of the Hamiltonian.

VI.6 Variational Integrators

A third approach to symplectic integrators comes from using discretized versions of Hamilton’s principle, which determines the equations of motion from a variational problem. This route has been taken by Suris (1990), MacKay (1992) and in a series of papers by Marsden and coauthors, see the review by Marsden & West (2001) and references therein. Basic theoretical properties were formulated by Maeda (1980,1982) and Veselov (1988,1991) in a non-numerical context.

VI.6.1 Hamilton’s Principle

Ours, according to Leibniz, is the best of all possible worlds, and the laws of nature can therefore be described in terms of extremal principles.

(C.L. Siegel & J.K. Moser 1971, p. 1)

Man scheint dies Princip früher ... unbemerkt gelassen zu haben.
Hamilton ist der erste, der von diesem Princip ausgegangen ist.

(C.G.J. Jacobi 1842, p. 58)

Hamilton gave an improved mathematical formulation of a principle which was well established by the fundamental investigations of Euler and Lagrange; the integration process employed by him was likewise known to Lagrange. The name “Hamilton’s principle”, coined by Jacobi, was not adopted by the scientists of the last century. It came into use, however, through the textbooks of more recent date.

(C. Lanczos 1949, p. 114)

Lagrange’s equations of motion (1.4) can be viewed as the Euler–Lagrange equations for the variational problem of extremizing the *action integral*

$$\mathcal{S}(q) = \int_{t_0}^{t_1} L(q(t), \dot{q}(t)) dt \quad (6.1)$$

among all curves $q(t)$ that connect two given points q_0 and q_1 :

$$q(t_0) = q_0, \quad q(t_1) = q_1. \quad (6.2)$$

In fact, assuming $q(t)$ to be extremal and considering a variation $q(t) + \varepsilon \delta q(t)$ with the same end-points, i.e., with $\delta q(t_0) = \delta q(t_1) = 0$, gives, using a partial integration,

$$0 = \left. \frac{d}{d\varepsilon} \right|_{\varepsilon=0} \mathcal{S}(q + \varepsilon \delta q) = \int_{t_0}^{t_1} \left(\frac{\partial L}{\partial q} \delta q + \frac{\partial L}{\partial \dot{q}} \delta \dot{q} \right) dt = \int_{t_0}^{t_1} \left(\frac{\partial L}{\partial q} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} \right) \delta q dt,$$

which leads to (1.4). The principle that the motion extremizes the action integral is known as *Hamilton’s principle*.

We now consider the action integral as a function of (q_0, q_1) , for the solution $q(t)$ of the Euler–Lagrange equations (1.4) with these boundary values (this exists uniquely locally at least if q_0, q_1 are sufficiently close),

$$S(q_0, q_1) = \int_{t_0}^{t_1} L(q(t), \dot{q}(t)) dt. \quad (6.3)$$

The partial derivative of S with respect to q_0 is, again using partial integration,

$$\begin{aligned} \frac{\partial S}{\partial q_0} &= \int_{t_0}^{t_1} \left(\frac{\partial L}{\partial q} \frac{\partial q}{\partial q_0} + \frac{\partial L}{\partial \dot{q}} \frac{\partial \dot{q}}{\partial q_0} \right) dt \\ &= \left. \frac{\partial L}{\partial \dot{q}} \frac{\partial q}{\partial q_0} \right|_{t_0}^{t_1} + \int_{t_0}^{t_1} \left(\frac{\partial L}{\partial q} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} \right) \frac{\partial q}{\partial q_0} dt = - \frac{\partial L}{\partial \dot{q}}(q_0, \dot{q}_0) \end{aligned}$$

with $\dot{q}_0 = \dot{q}(t_0)$, where the last equality follows from (1.4) and (6.2). In view of the definition (1.5) of the conjugate momenta, $p = \partial L / \partial \dot{q}$, the last term is simply $-p_0$. Computing $\partial S / \partial q_1 = p_1$ in the same way, we thus obtain for the differential of S

$$dS = \frac{\partial S}{\partial q_1} dq_1 + \frac{\partial S}{\partial q_0} dq_0 = p_1 dq_1 - p_0 dq_0 \quad (6.4)$$

which is the basic formula for symplecticity generating functions (see (5.1) above), obtained here by working with the Lagrangian formalism.

VI.6.2 Discretization of Hamilton's Principle

Discrete-time versions of Hamilton's principle are of mathematical interest in their own right, see Maeda (1980,1982), Veselov (1991) and references therein. Here they are considered with the aim of deriving or understanding numerical approximation schemes. The discretized Hamilton principle consists of extremizing, for given q_0 and q_N , the sum

$$\mathcal{S}_h(\{q_n\}_0^N) = \sum_{n=0}^{N-1} L_h(q_n, q_{n+1}) . \quad (6.5)$$

We think of the *discrete Lagrangian* L_h as an approximation

$$L_h(q_n, q_{n+1}) \approx \int_{t_n}^{t_{n+1}} L(q(t), \dot{q}(t)) dt , \quad (6.6)$$

where $q(t)$ is the solution of the Euler–Lagrange equations (1.4) with boundary values $q(t_n) = q_n$, $q(t_{n+1}) = q_{n+1}$. If equality holds in (6.6), then it is clear from the continuous Hamilton principle that the exact solution values $\{q(t_n)\}$ of the Euler–Lagrange equations (1.4) extremize the action sum \mathcal{S}_h . Before we turn to concrete examples of approximations L_h , we continue with the general theory which is analogous to the continuous case.

The requirement $\partial \mathcal{S}_h / \partial q_n = 0$ for an extremum yields the *discrete Euler–Lagrange equations*

$$\frac{\partial L_h}{\partial y}(q_{n-1}, q_n) + \frac{\partial L_h}{\partial x}(q_n, q_{n+1}) = 0 \quad (6.7)$$

for $n = 1, \dots, N-1$, where the partial derivatives refer to $L_h = L_h(x, y)$. This gives a three-term difference scheme for determining q_1, \dots, q_{N-1} .

We now set

$$S_h(q_0, q_N) = \sum_{n=0}^{N-1} L_h(q_n, q_{n+1})$$

where $\{q_n\}$ is a solution of the discrete Euler–Lagrange equations (6.7) with the boundary values q_0 and q_N . With (6.7) the partial derivatives reduce to

$$\frac{\partial S_h}{\partial q_0} = \frac{\partial L_h}{\partial x}(q_0, q_1), \quad \frac{\partial S_h}{\partial q_N} = \frac{\partial L_h}{\partial y}(q_{N-1}, q_N) .$$

We introduce the *discrete momenta* via a discrete Legendre transformation,

$$p_n = -\frac{\partial L_h}{\partial x}(q_n, q_{n+1}) . \quad (6.8)$$

The above formula and (6.7) for $n = N$ then yield

$$dS_h = p_N dq_N - p_0 dq_0 . \quad (6.9)$$

If (6.8) defines a bijection between p_n and q_{n+1} for given q_n , then we obtain a one-step method $\Phi_h : (p_n, q_n) \mapsto (p_{n+1}, q_{n+1})$ by composing the inverse discrete Legendre transform, a step with the discrete Euler–Lagrange equations, and the discrete Legendre transformation as shown in the diagram:

$$\begin{array}{ccc}
 & (6.7) & \\
 (q_n, q_{n+1}) & \longrightarrow & (q_{n+1}, q_{n+2}) \\
 (6.8) \uparrow & & \downarrow (6.8) \\
 (p_n, q_n) & & (p_{n+1}, q_{n+1})
 \end{array}$$

The method is symplectic by (6.9) and Theorem 5.1. A short-cut in the computation is obtained by noting that (6.7) and (6.8) (for $n + 1$ instead of n) imply

$$p_{n+1} = \frac{\partial L_h}{\partial y}(q_n, q_{n+1}), \quad (6.10)$$

which yields the scheme

$$(p_n, q_n) \xrightarrow{(6.8)} (q_n, q_{n+1}) \xrightarrow{(6.10)} (p_{n+1}, q_{n+1}).$$

Let us summarize these considerations, which can be found in Maeda (1980), Suris (1990), Veselov (1991) and MacKay (1992).

Theorem 6.1. *The discrete Hamilton principle for (6.5) gives the discrete Euler–Lagrange equations (6.7) and the symplectic method*

$$p_n = -\frac{\partial L_h}{\partial x}(q_n, q_{n+1}), \quad p_{n+1} = \frac{\partial L_h}{\partial y}(q_n, q_{n+1}). \quad (6.11)$$

These formulas also show that L_h is a generating function (5.4) for the symplectic map $(p_n, q_n) \mapsto (p_{n+1}, q_{n+1})$. Conversely, since every symplectic method has a generating function (5.4), it can be interpreted as resulting from Hamilton’s principle with the generating function (5.4) as the discrete Lagrangian. The classes of symplectic integrators and variational integrators are therefore identical.

We now turn to simple examples of variational integrators obtained by choosing a discrete Lagrangian L_h with (6.6).

Example 6.2 (MacKay 1992). Choose $L_h(q_n, q_{n+1})$ by approximating $q(t)$ of (6.6) as the linear interpolant of q_n and q_{n+1} and approximating the integral by the trapezoidal rule. This gives

$$L_h(q_n, q_{n+1}) = \frac{h}{2} L\left(q_n, \frac{q_{n+1} - q_n}{h}\right) + \frac{h}{2} L\left(q_{n+1}, \frac{q_{n+1} - q_n}{h}\right) \quad (6.12)$$

and hence the symplectic scheme, with $v_{n+1/2} = (q_{n+1} - q_n)/h$ for brevity,

$$\begin{aligned}
p_n &= \frac{1}{2} \frac{\partial L}{\partial \dot{q}}(q_n, v_{n+1/2}) + \frac{1}{2} \frac{\partial L}{\partial \dot{q}}(q_{n+1}, v_{n+1/2}) - \frac{h}{2} \frac{\partial L}{\partial q}(q_n, v_{n+1/2}) \\
p_{n+1} &= \frac{1}{2} \frac{\partial L}{\partial \dot{q}}(q_n, v_{n+1/2}) + \frac{1}{2} \frac{\partial L}{\partial \dot{q}}(q_{n+1}, v_{n+1/2}) + \frac{h}{2} \frac{\partial L}{\partial q}(q_{n+1}, v_{n+1/2}) .
\end{aligned}$$

For a mechanical Lagrangian $L(q, \dot{q}) = \frac{1}{2} \dot{q}^T M \dot{q} - U(q)$ this reduces to the Störmer–Verlet method

$$\begin{aligned}
Mv_{n+1/2} &= p_n + \frac{1}{2} h F_n \\
q_{n+1} &= q_n + h v_{n+1/2} \\
p_{n+1} &= Mv_{n+1/2} + \frac{1}{2} h F_{n+1}
\end{aligned}$$

where $F_n = -\nabla U(q_n)$. In this case, the discrete Euler–Lagrange equations (6.7) become the familiar second-difference formula $M(q_{n+1} - 2q_n + q_{n-1}) = h^2 F_n$.

Example 6.3 (Wendlandt & Marsden 1997). Approximating the integral in (6.6) instead by the midpoint rule gives

$$L_h(q_n, q_{n+1}) = hL\left(\frac{q_{n+1} + q_n}{2}, \frac{q_{n+1} - q_n}{h}\right). \quad (6.13)$$

This yields the symplectic scheme, with the abbreviations $q_{n+1/2} = (q_{n+1} + q_n)/2$ and $v_{n+1/2} = (q_{n+1} - q_n)/h$,

$$\begin{aligned}
p_n &= \frac{\partial L}{\partial \dot{q}}(q_{n+1/2}, v_{n+1/2}) - \frac{h}{2} \frac{\partial L}{\partial q}(q_{n+1/2}, v_{n+1/2}) \\
p_{n+1} &= \frac{\partial L}{\partial \dot{q}}(q_{n+1/2}, v_{n+1/2}) + \frac{h}{2} \frac{\partial L}{\partial q}(q_{n+1/2}, v_{n+1/2}) .
\end{aligned}$$

For $L(q, \dot{q}) = \frac{1}{2} \dot{q}^T M \dot{q} - U(q)$ this becomes the implicit midpoint rule

$$\begin{aligned}
Mv_{n+1/2} &= p_n + \frac{1}{2} h F_{n+1/2} \\
q_{n+1} &= q_n + h v_{n+1/2} \\
p_{n+1} &= Mv_{n+1/2} + \frac{1}{2} h F_{n+1/2}
\end{aligned}$$

with $F_{n+1/2} = -\nabla U(\frac{1}{2}(q_{n+1} + q_n))$.

VI.6.3 Symplectic Partitioned Runge–Kutta Methods Revisited

To obtain higher-order variational integrators, Marsden & West (2001) consider the discrete Lagrangian

$$L_h(q_0, q_1) = h \sum_{i=1}^s b_i L(u(c_i h), \dot{u}(c_i h)) \quad (6.14)$$

where $u(t)$ is the polynomial of degree s with $u(0) = q_0$, $u(h) = q_1$ which extremizes the right-hand side. They then show that the corresponding variational integrator can be realized as a partitioned Runge–Kutta method. We here consider the slightly more general case

$$L_h(q_0, q_1) = h \sum_{i=1}^s b_i L(Q_i, \dot{Q}_i) \quad (6.15)$$

where

$$Q_i = q_0 + h \sum_{j=1}^s a_{ij} \dot{Q}_j$$

and the \dot{Q}_i are chosen to extremize the above sum under the constraint

$$q_1 = q_0 + h \sum_{i=1}^s b_i \dot{Q}_i .$$

We assume that all the b_i are non-zero and that their sum equals 1. Note that (6.14) is the special case of (6.15) where the a_{ij} and b_i are integrals (II.1.10) of Lagrange polynomials as for collocation methods.

With a Lagrange multiplier $\lambda = (\lambda_1, \dots, \lambda_d)$ for the constraint, the extremality conditions obtained by differentiating (6.15) with respect to \dot{Q}_j for $j = 1, \dots, s$, read

$$\sum_{i=1}^s b_i \frac{\partial L}{\partial q}(Q_i, \dot{Q}_i) h a_{ij} + b_j \frac{\partial L}{\partial \dot{q}}(Q_j, \dot{Q}_j) = b_j \lambda .$$

With the notation

$$\dot{P}_i = \frac{\partial L}{\partial q}(Q_i, \dot{Q}_i) , \quad P_i = \frac{\partial L}{\partial \dot{q}}(Q_i, \dot{Q}_i) \quad (6.16)$$

this simplifies to

$$b_j P_j = b_j \lambda - h \sum_{i=1}^s b_i a_{ij} \dot{P}_i . \quad (6.17)$$

The symplectic method of Theorem 6.1 now becomes

$$\begin{aligned} p_0 &= -\frac{\partial L_h}{\partial x}(q_0, q_1) \\ &= -h \sum_{i=1}^s b_i \dot{P}_i \left(I + h \sum_{j=1}^s a_{ij} \frac{\partial \dot{Q}_j}{\partial q_0} \right) - h \sum_{j=1}^s b_j P_j \frac{\partial \dot{Q}_j}{\partial q_0} \\ &= -h \sum_{i=1}^s b_i \dot{P}_i + \lambda . \end{aligned}$$

In the last equality we use (6.17) and $h \sum_j b_j \partial \dot{Q}_j / \partial q_0 = -I$, which follows from differentiating the constraint. In the same way we obtain

$$p_1 = \frac{\partial L_h}{\partial y}(q_0, q_1) = \lambda .$$

Putting these formulas together, we see that (p_1, q_1) result from applying a partitioned Runge–Kutta method to the Lagrange equations (1.4) written as a differential-algebraic system

$$\dot{p} = \frac{\partial L}{\partial q}(q, \dot{q}) , \quad p = \frac{\partial L}{\partial \dot{q}}(q, \dot{q}) . \quad (6.18)$$

That is

$$\begin{aligned} p_1 &= p_0 + h \sum_{i=1}^s b_i \dot{P}_i , & q_1 &= q_0 + h \sum_{i=1}^s b_i \dot{Q}_i , \\ P_i &= p_0 + h \sum_{j=1}^s \hat{a}_{ij} \dot{P}_j , & Q_i &= q_0 + h \sum_{j=1}^s a_{ij} \dot{Q}_j , \end{aligned} \quad (6.19)$$

with $\hat{a}_{ij} = b_j - b_j a_{ji}/b_i$ so that the symplecticity condition (4.3) is fulfilled, and with $P_i, Q_i, \dot{P}_i, \dot{Q}_i$ related by (6.16). Since equations (6.16) are of the same form as (6.18), the proof of Theorem 1.3 shows that they are equivalent to

$$\dot{P}_i = -\frac{\partial H}{\partial q}(P_i, Q_i) , \quad \dot{Q}_i = \frac{\partial H}{\partial p}(P_i, Q_i) \quad (6.20)$$

with the Hamiltonian $H = p^T \dot{q} - L(q, \dot{q})$ of (1.6). We have thus proved the following, which is similar in spirit to a result of Suris (1990).

Theorem 6.4. *The variational integrator with the discrete Lagrangian (6.15) is equivalent to the symplectic partitioned Runge–Kutta method (6.19), (6.20) applied to the Hamiltonian system with the Hamiltonian (1.6).* \square

In particular, as noted by Marsden & West (2001), choosing Gaussian quadrature in (6.14) gives the Gauss collocation method applied to the Hamiltonian system, while Lobatto quadrature gives the Lobatto IIIA - IIIB pair.

VI.6.4 Noether's Theorem

... enthält Satz I alle in Mechanik u.s.w. bekannten Sätze über erste Integrale. (E. Noether 1918)

We now return to the subject of Chap. IV, i.e., the existence of first integrals, but here in the context of Hamiltonian systems. E. Noether found the surprising result that continuous *symmetries* in the Lagrangian lead to such first integrals. We give in the following a version of her “Satz I”, specialized to our needs, with a particularly short proof.

Theorem 6.5 (Noether 1918). *Consider a system with Hamiltonian $H(p, q)$ and Lagrangian $L(q, \dot{q})$. Suppose $\{g_s : s \in \mathbb{R}\}$ is a one-parameter group of transformations ($g_s \circ g_r = g_{s+r}$) which leaves the Lagrangian invariant:*

$$L(g_s(q), g'_s(q)\dot{q}) = L(q, \dot{q}) \quad \text{for all } s \text{ and all } (q, \dot{q}). \quad (6.21)$$

Let $a(q) = (d/ds)|_{s=0} g_s(q)$ be defined as the vector field with flow $g_s(q)$. Then

$$I(p, q) = p^T a(q) \quad (6.22)$$

is a first integral of the Hamiltonian system.

Example 6.6. Let G be a matrix Lie group with Lie algebra \mathfrak{g} (see Sect. IV.6). Suppose $L(Q\dot{q}, Q\dot{q}) = L(q, \dot{q})$ for all $Q \in G$. Then $p^T A q$ is a first integral for every $A \in \mathfrak{g}$. (Take $g_s(q) = \exp(sA)q$.) For example, $G = SO(n)$ yields conservation of angular momentum.

We prove Theorem 6.5 by using the discrete analogue, which reads as follows.

Theorem 6.7. Suppose the one-parameter group of transformations $\{g_s : s \in \mathbb{R}\}$ leaves the discrete Lagrangian $L_h(q_0, q_1)$ invariant:

$$L_h(g_s(q_0), g_s(q_1)) = L_h(q_0, q_1) \quad \text{for all } s \text{ and all } (q_0, q_1). \quad (6.23)$$

Then (6.22) is a first integral of the method (6.11), i.e., $p_{n+1}^T a(q_{n+1}) = p_n^T a(q_n)$.

Proof. Differentiating (6.23) with respect to s gives

$$0 = \frac{d}{ds} \Big|_{s=0} L_h(g_s(q_0), g_s(q_1)) = \frac{\partial L_h}{\partial x}(q_0, q_1) a(q_0) + \frac{\partial L_h}{\partial y}(q_0, q_1) a(q_1).$$

By (6.11) this becomes $0 = -p_0^T a(q_0) + p_1^T a(q_1)$. \square

Theorem 6.5 now follows by choosing $L_h = S$ of (6.3) and noting (6.4) and

$$\begin{aligned} S(q(t_0), q(t_1)) &= \int_{t_0}^{t_1} L(q(t), \dot{q}(t)) dt \\ &= \int_{t_0}^{t_1} L\left(g_s(q(t)), \frac{d}{dt} g_s(q(t))\right) dt = S\left(g_s(q(t_0)), g_s(q(t_1))\right). \end{aligned}$$

Theorem 6.7 has the appearance of giving a rich source of first integrals for symplectic methods. However, it must be noted that, unlike the case of the exact flow map in the above formula, the invariance (6.21) of the Lagrangian L does not in general imply the invariance (6.23) of the discrete Lagrangian L_h of the numerical method. A noteworthy exception arises for linear transformations g_s as in Example 6.6, for which Theorem 6.7 yields the conservation of quadratic first integrals $p^T A q$, such as angular momentum, by symplectic partitioned Runge–Kutta methods – a property we already know from Theorem IV.2.4. For Hamiltonian systems with an associated Lagrangian $L(q, \dot{q}) = \frac{1}{2} \dot{q}^T M \dot{q} - U(q)$, all first integrals originating from Noether's Theorem are quadratic (see Exercise 13).

VI.7 Characterization of Symplectic Methods

Up to now in this chapter, we have presented sufficient conditions for the symplecticity of numerical integrators (usually in terms of certain coefficients). Here, we will prove *necessary* conditions for symplecticity, i.e., answer the question as to which methods are *not* symplectic. It will turn out that the sufficient conditions of Sect. VI.4, under an irreducibility condition on the method, are also necessary. The main tool is the Taylor series expansion of the numerical flow $y_0 \mapsto \Phi_h(y_0)$, which we assume to be a B-series (or a P-series).

VI.7.1 B-Series Methods Conserving Quadratic First Integrals

The numerical solution of a Runge–Kutta method (II.1.4) can be written as a B-series

$$y_1 = B(a, y_0) = y_0 + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau) F(\tau)(y_0) \quad (7.1)$$

with coefficients $a(\tau)$ given by

$$a(\tau) = \sum_{i=1}^s b_i \mathbf{g}_i(\tau) \quad \text{for } \tau \in T \quad (7.2)$$

(see (III.1.16) and Sect. III.1.2). Our aim is to express the sufficient condition for the exact conservation of quadratic first integrals (which is the same as for symplecticity) in terms of the coefficients $a(\tau)$. For this we multiply (4.2) by $\mathbf{g}_i(u) \cdot \mathbf{g}_j(v)$ (where $u = [u_1, \dots, u_m]$ and $v = [v_1, \dots, v_l]$ are trees in T) and we sum over all i and j . Using (III.1.13) and the recursion (III.1.15) this yields

$$\sum_{i=1}^s b_i \mathbf{g}_i(u \circ v) + \sum_{j=1}^s b_j \mathbf{g}_j(v \circ u) = \left(\sum_{i=1}^s b_i \mathbf{g}_i(u) \right) \left(\sum_{j=1}^s b_j \mathbf{g}_j(v) \right),$$

where we have used the Butcher product (see, e.g., Butcher (1987), Sect. 143)

$$u \circ v = [u_1, \dots, u_m, v], \quad v \circ u = [v_1, \dots, v_l, u] \quad (7.3)$$

(compare also Definition III.3.7 and Fig. 7.1 below). Because of (7.2), this implies

$$a(u \circ v) + a(v \circ u) = a(u) \cdot a(v) \quad \text{for } u, v \in T. \quad (7.4)$$

We now forget that the B-series (7.1) has been obtained from a Runge–Kutta method, and we ask the following question: is the condition (7.4) sufficient for a B-series method defined by (7.1) to conserve exactly quadratic first integrals (and to be symplectic)? The next theorem shows that this is indeed true, and we shall see later that condition (7.4) is also necessary (cf. Chartier, Faou & Murua 2005).

Theorem 7.1. Consider a B-series method $\Phi_h(y) = B(a, y)$ and problems $\dot{y} = f(y)$ having $Q(y) = y^T C y$ (with symmetric matrix C) as first integral.

If the coefficients $a(\tau)$ satisfy (7.4), then the method exactly conserves $Q(y)$ and it is symplectic.

Proof. a) Under the assumptions of the theorem we shall prove in part (c) that

$$B(a, y)^T C B(a, y) = y^T C y + \sum_{u, v \in T} \frac{h^{|u|+|v|}}{\sigma(u)\sigma(v)} m(u, v) F(u)(y)^T C F(v)(y) \quad (7.5)$$

with $m(u, v) = a(u) \cdot a(v) - a(u \circ v) - a(v \circ u)$. Condition (7.4) is equivalent to $m(u, v) = 0$ and thus implies the exact conservation of $Q(y) = y^T C y$.

To prove symplecticity of the method it is sufficient to show that the diagram of Lemma 4.1 commutes for general B-series methods. This is seen by differentiating the elementary differentials and by comparing them with those for the augmented system (Exercise 8). Symplecticity of the method thus follows as in Sect. VI.4.1 from the fact that the symplecticity relation is a quadratic first integral of the augmented system.

b) Since $Q(y) = y^T C y$ is a first integral of $\dot{y} = f(y)$, we have $y^T C f(y) = 0$ for all y . Differentiating m times this relation with respect to y yields

$$\sum_{j=1}^m k_j^T C f^{(m-1)}(y)(k_1, \dots, k_{j-1}, k_{j+1}, \dots, k_m) + y^T C f^{(m)}(y)(k_1, \dots, k_m) = 0.$$

Putting $k_j = F(\tau_j)(y)$ we obtain the formula

$$y^T C F([\tau_1, \dots, \tau_m])(y) = - \sum_{j=1}^m F(\tau_j)(y)^T C F([\tau_1, \dots, \tau_{j-1}, \tau_{j+1}, \dots, \tau_m])(y),$$

which can also be written in the form

$$y^T C \frac{F(\tau)(y)}{\sigma(\tau)} = - \sum_{u, v \in T, v \circ u = \tau} \frac{F(u)(y)^T}{\sigma(u)} C \frac{F(v)(y)}{\sigma(v)}. \quad (7.6)$$

c) With (7.1) the expression $y_1^T C y_1$ becomes

$$\begin{aligned} B(a, y)^T C B(a, y) &= y^T C y + 2y^T C \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} a(\tau) F(\tau)(y) \\ &\quad + \sum_{u, v \in T} \frac{h^{|u|+|v|}}{\sigma(u)\sigma(v)} a(u) a(v) F(u)(y)^T C F(v)(y). \end{aligned}$$

Since C is symmetric, formula (7.6) remains true if we sum over trees u, v such that $u \circ v = \tau$. Inserting both formulas into the sum over τ leads directly to (7.5). \square

Extension to P-Series. All the previous results can be extended to partitioned methods. To find the correct conditions on the coefficients of the P-series, we use the fact that the numerical solution of a partitioned Runge–Kutta method (II.2.2) is a P-series

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \begin{pmatrix} P_p(a, (p_0, q_0)) \\ P_q(a, (p_0, q_0)) \end{pmatrix} = \begin{pmatrix} p_0 \\ q_0 \end{pmatrix} + \begin{pmatrix} \sum_{u \in TP_p} \frac{h|u|}{\sigma(u)} a(u) F(u)(p_0, q_0) \\ \sum_{v \in TP_q} \frac{h|v|}{\sigma(v)} a(v) F(v)(p_0, q_0) \end{pmatrix} \quad (7.7)$$

with coefficients $a(\tau)$ given by

$$a(\tau) = \begin{cases} \sum_{i=1}^s b_i \phi_i(\tau) & \text{for } \tau \in TP_p \\ \sum_{i=1}^s \hat{b}_i \phi_i(\tau) & \text{for } \tau \in TP_q \end{cases} \quad (7.8)$$

(see Theorem III.2.4). We assume here that the elementary differentials $F(\tau)(p, q)$ originate from a partitioned system

$$\dot{p} = f_1(p, q), \quad \dot{q} = f_2(p, q), \quad (7.9)$$

such as the Hamiltonian system (1.7). This time we multiply (4.3) by $\phi_i(u) \cdot \phi_j(v)$ (where $u = [u_1, \dots, u_m]_p \in TP_p$ and $v = [v_1, \dots, v_l]_q \in TP_q$) and we sum over all i and j . Using the recursion (III.2.7) this yields

$$\sum_{i=1}^s b_i \phi_i(u \circ v) + \sum_{j=1}^s \hat{b}_j \phi_j(v \circ u) = \left(\sum_{i=1}^s b_i \phi_i(u) \right) \left(\sum_{j=1}^s \hat{b}_j \phi_j(v) \right), \quad (7.10)$$

where $u \circ v = [u_1, \dots, u_m, v]_p$ and $v \circ u = [v_1, \dots, v_l, u]_q$. Because of (7.8), this implies the relation

$$a(u \circ v) + a(v \circ u) = a(u) \cdot a(v) \quad \text{for } u \in TP_p, v \in TP_q. \quad (7.11)$$

Since $\phi_i(\tau)$ is independent of the colour of the root of τ , condition (4.4) implies

$$a(\tau) \text{ is independent of the colour of the root of } \tau. \quad (7.12)$$

Theorem 7.2. Consider a P-series method $(p_1, q_1) = \Phi_h(p_0, q_0)$ given by (7.7), and a problem (7.9) having $Q(p, q) = p^T E q$ as first integral.

i) If the coefficients $a(\tau)$ satisfy (7.11) and (7.12), the method exactly conserves $Q(p, q)$ and it is symplectic for general Hamiltonian systems (1.7).

ii) If the coefficients $a(\tau)$ satisfy only (7.11), the method exactly conserves $Q(p, q)$ for problems of the form $\dot{p} = f_1(q)$, $\dot{q} = f_2(p)$, and it is symplectic for separable Hamiltonian systems where $H(p, q) = T(p) + U(q)$.

Proof. This is very similar to that of Theorem 7.1. If $Q(p, q) = p^T E q$ is a first integral of (7.9), we have $f_1(p, q)^T E q + p^T E f_2(p, q) = 0$ for all p and q . Differentiating m times with respect to p and n times with respect to q yields

$$\begin{aligned}
0 &= D_p^m D_q^n f_1(p, q) (k_1, \dots, k_m, \ell_1, \dots, \ell_n)^T E q \\
&+ p^T E D_p^m D_q^n f_2(p, q) (k_1, \dots, k_m, \ell_1, \dots, \ell_n) \\
&+ \sum_{j=1}^n D_p^m D_q^{n-1} f_1(p, q) (k_1, \dots, k_m, \ell_1, \dots, \ell_{j-1}, \ell_{j+1}, \dots, \ell_n)^T E \ell_j \\
&+ \sum_{i=1}^m k_i^T E D_p^{m-1} D_q^n f_2(p, q) (k_1, \dots, k_{i-1}, k_{i+1}, \dots, k_m, \ell_1, \dots, \ell_n).
\end{aligned} \tag{7.13}$$

Putting $k_i = F(u_i)(p, q)$ with $u_i \in TP_p$, $\ell_j = F(v_j)(p, q)$ with $v_j \in TP_q$, $\tau_p = [u_1, \dots, u_m, v_1, \dots, v_n]_p$ and $\tau_q = [u_1, \dots, u_m, v_1, \dots, v_n]_q$, we obtain as in part (b) of the proof of Theorem 7.1 that

$$\begin{aligned}
&\frac{F(\tau_p)(p, q)^T}{\sigma(\tau_p)} E q + p^T E \frac{F(\tau_q)(p, q)}{\sigma(\tau_q)} \\
&= \sum_{u \circ v = \tau_p} \frac{F(u)(p, q)^T}{\sigma(u)} E \frac{F(v)(p, q)}{\sigma(v)} + \sum_{v \circ u = \tau_q} \frac{F(u)(p, q)^T}{\sigma(u)} E \frac{F(v)(p, q)}{\sigma(v)},
\end{aligned} \tag{7.14}$$

where the sums are over $u \in TP_p$ and $v \in TP_q$.

With (7.7) the expression $p_1^T E q_1$ becomes

$$\begin{aligned}
P_p(a, (p, q))^T E P_q(a, (p, q)) &= p^T E q \\
&+ \sum_{u \in TP_p} \frac{h^{|u|}}{\sigma(u)} a(u) F(u)(p, q)^T E q + p^T E \sum_{v \in TP_q} \frac{h^{|v|}}{\sigma(v)} a(v) F(v)(p, q) \\
&+ \sum_{u \in TP_p, v \in TP_q} \frac{h^{|u|+|v|}}{\sigma(u)\sigma(v)} a(u)a(v) F(u)(p, q)^T E F(v)(p, q).
\end{aligned} \tag{7.15}$$

Condition (7.12) implies that $a(\tau_p) = a(\tau_q)$ for the trees in (7.14). Since also $|\tau_p| = |\tau_q|$ and $\sigma(\tau_p) = \sigma(\tau_q)$, two corresponding terms in the sums of the second line in (7.15) can be jointly replaced by the use of (7.14). As in part (c) of the proof of Theorem 7.1 this together with (7.11) then yields

$$P_p(a, (p, q))^T E P_q(a, (p, q)) = p^T E q,$$

which proves the conservation of quadratic first integrals $p^T E q$. Symplecticity follows as before, because the diagram of Lemma 4.1 also commutes for general P-series methods.

For the proof of statement (ii) we notice that $f_1(q)^T E q + p^T E f_2(p) = 0$ implies that $f_1(q)^T E q = 0$ and $p^T E f_2(p) = 0$ vanish separately. Instead of (7.14) we thus have two identities: the term $F(\tau_p)(p, q)^T E q / \sigma(\tau_p)$ becomes equal to the first sum in (7.14), and $p^T E F(\tau_q)(p, q) / \sigma(\tau_q)$ to the second sum. Consequently, the previous argumentation can be applied without the condition $a(\tau_p) = a(\tau_q)$. \square

Second Order Differential Equations. We next consider partitioned systems of the particular form

$$\dot{p} = f_1(q), \quad \dot{q} = Cp + c, \quad (7.16)$$

where C is a matrix and c a vector. Since problems of this type are second order differential equations $\ddot{q} = Cf_1(q)$, partitioned Runge–Kutta methods become equivalent to Nyström methods (see Sects. II.2.3 and IV.2.3).

An important special case are Hamiltonian systems

$$\dot{p} = -\nabla U(q), \quad \dot{q} = Cp + c \quad (7.17)$$

(or, equivalently, $\ddot{q} = -C\nabla U(q)$). They correspond to Hamiltonian functions

$$H(p, q) = \frac{1}{2} p^T Cp + c^T p + U(q), \quad (7.18)$$

where the kinetic energy is at most quadratic in p (here, C is usually symmetric).

In a P-series representation of the numerical solution, many elementary differentials vanish identically. Only those trees have to be considered, whose neighbouring vertices have different colour (the problem is separable) and whose white vertices have at most one son⁶ (second component is linear). We denote this set of trees by

$$TN_p = \left\{ \tau \in TP_p \mid \begin{array}{l} \text{neighbouring vertices of } \tau \text{ have different colour} \\ \text{white vertices of } \tau \text{ have at most one son} \end{array} \right\}, \quad (7.19)$$

and we let TN_q be the corresponding subset of TP_q .

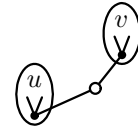
The same procedure as for partitioned methods permits us to write the symplecticity condition of Theorem 4.8 in terms of the coefficients $a(\tau)$ of the P-series. Assuming $a(\bullet) = a(\circ) = 1$, the two conditions of (4.5) lead to

$$a(\circ \circ u) + a(u \circ \circ) = a(u)a(\circ) \quad \text{for } u \in TN_p \quad (7.20)$$

$$a(u)a(\circ \circ v) - a(u \circ \circ v) = a(\circ \circ u)a(v) - a(v \circ \circ u) \quad \text{for } u, v \in TN_p \quad (7.21)$$

where we use the abbreviating notation

$$u \circ \circ v = u \circ (\circ \circ v) = [u_1, \dots, u_m, [v]_q]_p \quad (7.22)$$



if $u = [u_1, \dots, u_m]_p$. Notice that for $u, v \in TN_p$, the trees $u \circ \circ$, $u \circ \circ v$ and $v \circ \circ u$ are in TN_p , and $\circ \circ u$ is in TN_q .

Theorem 7.3. Consider a P-series method (7.7) for differential equations (7.16) having $Q(p, q) = p^T E q$ as first integral.

If the coefficients $a(\tau)$ satisfy (7.20) and (7.21), the method exactly conserves $Q(p, q)$ and it is symplectic for Hamiltonian systems with $H(p, q)$ of the form (7.18).

⁶ Attention: with respect to (III.2.10) the vertices have opposite colour, because the linear dependence is in the second component in (7.17) whereas it is in the first component in (III.2.9).

Proof. Since the elementary differentials $F(\tau)(p, q)$ vanish identically for $\tau \notin TN_p \cup TN_q$, we can arbitrarily define $a(\tau)$ for trees outside $TN_p \cup TN_q$ without changing the method (throughout this proof we implicitly assume that for the considered trees neighbouring vertices have different colour). We shall do this in such a way that (7.11) holds.

Consider first the tree $u \circ \circ v$. There is exactly one vertex between the roots of u and v . Making this vertex to the root gives the tree $[u, v]_q$ which is not in TN_q . We define for $u, v \in TN_p$

$$a([u, v]_q) := a(u)a(\circ \circ v) - a(u \circ \circ v).$$

Condition (7.21) shows that $a([u, v]_q)$ is independent of permuting u and v and is thus well-defined. For trees that are neither in $TN_p \cup TN_q$ nor of the form $[u, v]_q$ with $u, v \in TN_p$ we let $a(\tau) = 0$. This extension of $a(\tau)$ implies that condition (7.11) holds for all trees, and part (ii) of Theorem 7.2 yields the statement. Notice that for problems $\dot{p} = f_1(q)$, $\dot{q} = f_2(p)$ only trees, for which neighbouring vertices have different colour, are relevant. \square

VI.7.2 Characterization of Symplectic P-Series (and B-Series)

A characterization of symplectic B-series was first obtained by Calvo & Sanz-Serna (1994). We also consider P-series with various important special situations.

Theorem 7.4. *Consider a P-series method (7.7) applied to a general partitioned differential equation (7.9). Equivalent are:*

- 1) *the coefficients $a(\tau)$ satisfy (7.11) and (7.12),*
- 2) *quadratic first integrals of the form $Q(p, q) = p^T E q$ are exactly conserved,*
- 3) *the method is symplectic for general Hamiltonian systems (1.7).*

Proof. The implication (1) \Rightarrow (2) follows from part (i) of Theorem 7.2, (2) \Rightarrow (3) is a consequence of the fact that the symplecticity condition is a quadratic first integral of the variational equation (see the proof of Theorem 7.2). The remaining implication (3) \Rightarrow (1) will be proved in the following two steps.

a) We fix two trees $u \in TP_p$ and $v \in TP_q$, and we construct a (polynomial) Hamiltonian such that the transformation (7.7) satisfies

$$\left(\frac{\partial(p_1, q_1)}{\partial p_0^1} \right)^T J \left(\frac{\partial(p_1, q_1)}{\partial q_0^2} \right) = C \left(a(u \circ v) + a(v \circ u) - a(u) \cdot a(v) \right) \quad (7.23)$$

with $C \neq 0$ (here, p_0^1 denotes the first component of p_0 , and q_0^2 the second component of q_0). The symplecticity of (7.7) implies that the expression in (7.23) vanishes, so that condition (7.11) has to be satisfied.

For given $u \in TP_p$ and $v \in TP_q$ we define the Hamiltonian as follows: to the branches of $u \circ v$ we attach the numbers $3, \dots, |u| + |v| + 1$ such that the branch between the roots of u and v is labelled by 3. Then, the Hamiltonian is a sum of as many terms as vertices in the tree. The summand corresponding to a vertex is a

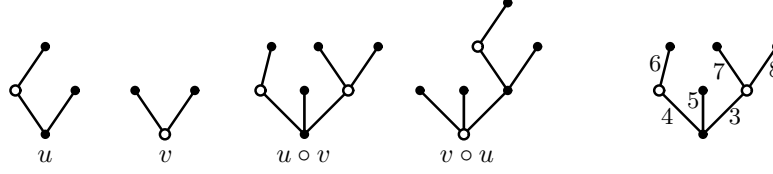


Fig. 7.1. Illustration of the Hamiltonian (7.24)

product containing the factor p^j (resp. q^j) if an upward leaving branch “ j ” is directly connected with a black (resp. white) vertex, and the factor q^i (resp. p^i) if the vertex itself is black (resp. white) and the downward leaving branch has label “ i ”. Finally, the factors q^2 and p^1 are included in the terms corresponding to the roots of u and v , respectively. For the example of Fig. 7.1 we have

$$H(p, q) = q^2 q^3 q^4 p^5 + p^1 p^3 p^7 p^8 + p^4 p^6 + q^5 + q^6 + q^7 + q^8. \quad (7.24)$$

The components $F^i(\tau)(p, q)$ of the elementary differentials corresponding to the Hamiltonian system (with the Hamiltonian constructed above) satisfy

$$\begin{aligned} F^2(u \circ v)(p, q) &= (-1)^{\delta(u \circ v)} \sigma(u \circ v) \cdot p^1, \\ F^1(v \circ u)(p, q) &= (-1)^{\delta(v \circ u)} \sigma(v \circ u) \cdot q^2, \\ F^3(u)(p, q) &= (-1)^{\delta(u)} \sigma(u) \cdot q^2, \\ F^3(v)(p, q) &= (-1)^{\delta(v)} \sigma(v) \cdot p^1, \end{aligned} \quad (7.25)$$

and for all other trees $\tau \in TP$ and components i we have

$$\frac{\partial F^i(\tau)}{\partial p^1}(0, 0) = \frac{\partial F^i(\tau)}{\partial q^2}(0, 0) = 0.$$

In (7.25), $\delta(\tau)$ counts the number of black vertices of τ , and the *symmetry coefficient* $\sigma(\tau)$ is that of (III.2.3). For example, $\sigma(u) = 1$ and $\sigma(v) = 2$ for the trees of Fig. 7.1. The verification of (7.25) is straightforward. The coefficient $(-1)^{\delta(\tau)}$ is due to the minus sign in the first part of the Hamiltonian system (1.7), and the symmetry coefficient $\sigma(\tau)$ appears in exactly the same way as in the multidimensional Taylor formula. Due to the zero initial values, no elementary differential other than those of (7.25) give rise to non-vanishing expressions in (7.23). Consider for example the second component of $F(\tau)(p, q)$ for a tree $\tau \in TP_p$. Since we are concerned with the Hamiltonian system (1.7), this expression starts with a derivative of H_{q^2} . Therefore, it contributes to (7.23) at $p_0 = q_0 = 0$ only if it contains the factor $H_{q^2 q^3 q^4 p^5}$ (for the example of Fig. 7.1). This in turn implies the presence of factors $H_{p^3 \dots}$, $H_{p^4 \dots}$ and $H_{q^5 \dots}$. Continuing this line of reasoning, we find that $F^2(\tau)(p, q)$ contributes to (7.23) at $p_0 = q_0 = 0$ only if $\tau = u \circ v$. With similar arguments we see that only the elementary differentials of (7.25) have to be considered. We now insert (7.25) into (7.7), and we compute its derivatives with respect to p^1 and q^2 . This then yields (7.23) with $C = (-1)^{\delta(u) + \delta(v)} h^{|u| + |v|}$, and completes the proof concerning condition (7.11).

b) The necessity of condition (7.12) is seen similarly. We fix a tree $\tau \in TP_p$ and we let $\bar{\tau} \in TP_q$ be the tree obtained from τ by changing the colour of the root. We then attach the numbers $3, \dots, |\tau| + 1$ to the branches of τ , and we define a Hamiltonian as above but, different from adding the factors q^2 and p^1 , we include the factor $p^1 q^2$ to the term corresponding to the root. For the tree $\tau = u$ of Fig. 7.1 this yields

$$H(p, q) = p^1 q^2 q^3 p^4 + p^3 p^5 + q^4 + q^5.$$

With this Hamiltonian we get

$$\begin{aligned} F^2(\tau)(p, q) &= (-1)^{\delta(\tau)} \sigma(\tau) \cdot p^1, \\ F^1(\bar{\tau})(p, q) &= (-1)^{\delta(\tau)} \sigma(\tau) \cdot q^2, \end{aligned}$$

and these are the only elementary differentials contributing to the left-hand expression of (7.23). We thus get

$$\left(\frac{\partial(p_1, q_1)}{\partial p_0^1} \right)^T J \left(\frac{\partial(p_1, q_1)}{\partial q_0^2} \right) = (-1)^{\delta(\tau)} h^{|\tau|} (a(\tau) - a(\bar{\tau})),$$

which completes the proof of Theorem 7.4. \square

Theorem 7.5. *Consider a P-series method (7.7) applied to a separable partitioned differential equation $\dot{p} = f_1(q)$, $\dot{q} = f_2(p)$. Equivalent are:*

- 1) *the coefficients $a(\tau)$ satisfy (7.11),*
- 2) *quadratic first integrals of the form $Q(p, q) = p^T E q$ are exactly conserved,*
- 3) *the method is symplectic for separable Hamiltonians $H(p, q) = T(p) + U(q)$.*

Proof. The implications (1) \Rightarrow (2) \Rightarrow (3) follow as before from part (ii) of Theorem 7.2. The remaining implication (3) \Rightarrow (1) is a consequence of the fact that the Hamiltonian constructed in part (a) of the proof of Theorem 7.4 is separable, when u and v have no neighbouring vertices of the same colour. \square

Theorem 7.6. *Consider a B-series method (7.1) for $\dot{y} = f(y)$. Equivalent are:*

- 1) *the coefficients $a(\tau)$ satisfy (7.4),*
- 2) *quadratic first integrals of the form $Q(y) = y^T C y$ are exactly conserved,*
- 3) *the method is symplectic for general Hamiltonian systems $\dot{y} = J^{-1} \nabla H(y)$.*

Proof. The implications (1) \Rightarrow (2) \Rightarrow (3) follow from Theorem 7.1. The remaining implication (3) \Rightarrow (1) follows from Theorem 7.4, because a B-series with coefficients $a(\tau)$, $\tau \in T$, applied to a partitioned differential equation, can always be interpreted as a P-series (Definition III.2.1), where $a(\tau) := a(\varphi(\tau))$ for $\tau \in TP$ and $\varphi : TP \rightarrow T$ is the mapping that forgets the colouring of the vertices. This follows from the fact that

$$\alpha(\tau) F(\tau)(y) = \left(\frac{\sum_{u \in TP_p, \varphi(u)=\tau} \alpha(u) F(u)(p, q)}{\sum_{v \in TP_q, \varphi(v)=\tau} \alpha(v) F(v)(p, q)} \right)$$

for $\tau \in T$, because $\alpha(u) \cdot \sigma(u) = \alpha(v) \cdot \sigma(v) = \mathbf{e}(\tau) \cdot |\tau|!$. Here, $y = (p, q)$, the elementary differentials $F(\tau)(y)$ are those of Definition III.1.2, whereas $F(u)(p, q)$ and $F(v)(p, q)$ are those of Table III.2.1. \square

Theorem 7.7. *Consider a P-series method (7.7) applied to the special partitioned system (7.16). Equivalent are:*

- 1) *the coefficients $a(\tau)$ satisfy (7.20) and (7.21),*
- 2) *quadratic first integrals of the form $Q(p, q) = p^T E q$ are exactly conserved,*
- 3) *the method is symplectic for Hamiltonian systems of the form (7.17).*

Proof. The implications (1) \Rightarrow (2) \Rightarrow (3) follow from Theorem 7.3. The remaining implication (3) \Rightarrow (1) can be seen as follows.

Condition (7.20) is a consequence of the the proof of Theorem 7.4, because for $u \in TN_p$ and $v = \circ$ the Hamiltonian constructed there is of the form (7.18).

To prove condition (7.21) we have to modify slightly the definition of $H(p, q)$. We take $u, v \in TN_p$ and define the polynomial Hamiltonian as follows: to the branches of $u \circ \circ v$ we attach the numbers $3, \dots, |u| + |v| + 2$. The Hamiltonian is then a sum of as many terms as vertices in the tree. The summands are defined as in the proof of Theorem 7.4 with the only exception that to the terms corresponding to the roots of u and v we include the factors q^2 and q^1 , respectively, instead of q^2 and p^1 . This gives a Hamiltonian of the form (7.18), for which the expression

$$\left(\frac{\partial(p_1, q_1)}{\partial q_0^1} \right)^T J \left(\frac{\partial(p_1, q_1)}{\partial q_0^2} \right) \quad (7.26)$$

becomes equal to

$$a(u)a(\circ \circ v) - a(u \circ \circ v) - a(\circ \circ u)a(v) + a(v \circ \circ u) \quad (7.27)$$

up to a nonzero constant. By symplecticity, (7.26) is zero so that also (7.27) has to vanish. This proves the validity of condition (7.21). \square

VI.7.3 Irreducible Runge–Kutta Methods

We are now able to study to what extent the conditions of Theorem 4.3 and Theorem 4.6 are also necessary for symplecticity. Consider first the 2-stage method

$$\begin{array}{c|cc} 1/2 & \alpha & 1/2 - \alpha \\ 1/2 & \beta & 1/2 - \beta \\ \hline & 1/2 & 1/2 \end{array}.$$

The solution of the corresponding Runge–Kutta system (II.1.4) is given by $k_1 = k_2 = k$, where $k = f(y_0 + k/2)$, and hence $y_1 = y_0 + hk$. Whatever the values of α and β are, the numerical solution of the Runge–Kutta method is identical to that of the implicit midpoint rule, so that it defines a symplectic transformation. However, the condition (4.2) is only satisfied for $\alpha = \beta = 1/4$.

Definition 7.8. Two stages i and j of a Runge–Kutta method (II.1.4) are said to be *equivalent for a class (\mathcal{P})* of initial value problems, if for every problem in (\mathcal{P}) and for every sufficiently small step size we have $k_i = k_j$ ($k_i = k_j$ and $\ell_i = \ell_j$ for partitioned Runge–Kutta methods (II.2.2)).

The method is called *irreducible for* (\mathcal{P}) if it does not have equivalent stages. It is called *irreducible* if it is irreducible for all sufficiently smooth initial value problems.

For a more amenable characterization of irreducible Runge–Kutta methods, we introduce an ordering on T (and on TP), and we consider the following $s \times \infty$ matrices

$\Phi_{\text{RK}} = (\phi(\tau); \tau \in T)$ with entries $\phi_i(\tau) = \mathbf{g}_i(\tau)$ given by (III.1.13),⁷
 $\Phi_{\text{PRK}} = (\phi(\tau); \tau \in TP_p) = (\phi(\tau); \tau \in TP_q)$ with entries $\phi_i(\tau)$ given by (III.2.7);
 observe that $\phi_i(\tau)$ does not depend on the colour of the root,
 $\Phi_{\text{PRK}}^* = (\phi(\tau); \tau \in TP_p^*) = (\phi(\tau); \tau \in TP_q^*)$ where TP_p^* (resp. TP_q^*) is the set of trees in TP_p (resp. TP_q) whose neighbouring vertices have different colours.

Lemma 7.9 (Hairer 1994). *A Runge–Kutta method is irreducible if and only if the matrix Φ_{RK} has full rank s .*

A partitioned Runge–Kutta method is irreducible if and only if the matrix Φ_{PRK} has full rank s .

A partitioned Runge–Kutta method is irreducible for separable problems $\dot{p} = f_1(q)$, $\dot{q} = f_2(p)$ if and only if the matrix Φ_{PRK}^ has full rank s .*

Proof. If the stages i and j are equivalent, it follows from the expansion

$$k_i = \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} \phi_i(\tau) F(\tau)(y_0)$$

(see the proof of Theorem III.1.4) and from the independency of the elementary differentials (Exercise III.3) that $\phi_i(\tau) = \phi_j(\tau)$ for all $\tau \in T$. Hence, the rows i and j of the matrix Φ_{RK} are identical. The analogous statement for partitioned Runge–Kutta methods follows from Theorem III.2.4 and Exercise III.6. This proves the sufficiency of the “full rank” condition.

We prove its necessity only for partitioned Runge–Kutta methods applied to separable problems (the other situations can be treated similarly). For separable problems, only trees in $TP_p^* \cup TP_q^*$ give rise to non-vanishing elementary differentials. Irreducibility therefore implies that for every pair (i, j) with $i \neq j$ there exists a tree $\tau \in TP_p^*$ such that $\phi_i(\tau) \neq \phi_j(\tau)$. Consequently, a certain finite linear combination of the columns of Φ_{PRK}^* has distinct elements, i.e., there exist vectors $\xi \in \mathbb{R}^\infty$ (only finitely many non zero elements) and $\eta \in \mathbb{R}^s$ with $\Phi_{\text{PRK}}^* \xi = \eta$ and $\eta_i \neq \eta_j$ for $i \neq j$. Due to the fact that $\phi_i([\tau_1, \dots, \tau_m]) = \phi_i([\tau_1]) \cdot \dots \cdot \phi_i([\tau_m])$, the componentwise product of two columns of Φ_{PRK}^* is again a column of Φ_{PRK}^* . Continuing this argumentation and observing that $(1, \dots, 1)^T$ is a column of Φ_{PRK}^* , we obtain a matrix X such that $\Phi_{\text{PRK}}^* X = (\eta_i^{j-1})_{i,j=1}^s$ is a Vandermonde matrix. Since the η_i are distinct, the matrix Φ_{PRK}^* has to be of full rank s . \square

⁷ In this section we let $\phi(\tau) \in \mathbb{R}^s$ denote the vector whose elements are $\phi_i(\tau)$, $i = 1, \dots, s$. This should not be mixed up with the value $\phi(\tau)$ of (III.1.16).

VI.7.4 Characterization of Irreducible Symplectic Methods

The necessity of the condition (4.2) for symplectic Runge–Kutta methods was first stated by Lasagni (1988). Abia & Sanz-Serna (1993) extended his proof to partitioned methods. We follow here the ideas of Hairer (1994).

Theorem 7.10. *An irreducible Runge–Kutta method (II.1.4) is symplectic if and only if the condition (4.2) holds.*

An irreducible partitioned Runge–Kutta method (II.2.2) is symplectic if and only if the conditions (4.3) and (4.4) hold.

A partitioned Runge–Kutta method, irreducible for separable problems, is symplectic for separable Hamiltonians $H(p, q) = T(p) + U(q)$ if and only if the condition (4.3) holds.

Proof. The “if” part of all three statements has been proved in Theorem 4.3 and Theorem 4.6. We prove the “only if” part for partitioned Runge–Kutta methods applied to general Hamiltonian systems (the other two statements can be obtained in the same way).

We consider the $s \times s$ matrix M with entries $m_{ij} = b_i \hat{a}_{ij} + \hat{b}_j a_{ji} - b_i \hat{b}_j$. The computation leading to formula (7.11) shows that for $u \in TP_p$ and $v \in TP_q$

$$\phi(u)^T M \phi(v) = a(u \circ v) + a(v \circ u) - a(u) \cdot a(v)$$

holds. Due to the symplecticity of the method, this expression vanishes and we obtain

$$\Phi_{\text{PRK}}^T M \Phi_{\text{PRK}} = 0,$$

where Φ_{PRK} is the matrix of Lemma 7.9. An application of this lemma then yields $M = 0$, which proves the necessity of (4.3).

For the vector d with components $d_i = b_i - \hat{b}_i$ we get $d^T \Phi_{\text{PRK}} = 0$, and we deduce from Lemma 7.9 that $d = 0$, so that (4.4) is also seen to be necessary. \square

VI.8 Conjugate Symplecticity

The symplecticity requirement may be too strong if we are interested in a correct long-time behaviour of a numerical integrator. Stoffer (1988) suggests considering methods that are not necessarily symplectic but conjugate to a symplectic method.

Definition 8.1. Two numerical methods Φ_h and Ψ_h are mutually *conjugate*, if there exists a global change of coordinates χ_h , such that

$$\Phi_h = \chi_h^{-1} \circ \Psi_h \circ \chi_h. \quad (8.1)$$

We assume that $\chi_h(y) = y + \mathcal{O}(h)$ uniformly for y varying in a compact set.

For a numerical solution $y_{n+1} = \Phi_h(y_n)$, lying in a compact subset of the phase space, the transformed values $z_n = \chi_h(y_n)$ constitute a numerical solution $z_{n+1} = \Psi_h(z_n)$ of the second method. Since $y_n - z_n = \mathcal{O}(h)$, both numerical solutions have the same long-time behaviour, independently of whether one method shares certain properties (e.g., symplecticity) with the other.

VI.8.1 Examples and Order Conditions

The most prominent pair of conjugate methods are the trapezoidal and midpoint rules. Their conjugacy has been originally exploited by Dahlquist (1975) in an investigation on nonlinear stability.

If we denote by Φ_h^E and Φ_h^I the explicit and implicit Euler methods, respectively, then the trapezoidal rule Φ_h^T and the implicit midpoint rule Φ_h^M can be written as

$$\Phi_h^T = \Phi_{h/2}^I \circ \Phi_{h/2}^E, \quad \Phi_h^M = \Phi_{h/2}^E \circ \Phi_{h/2}^I$$

(see Fig. 8.1). This shows $\Phi_h^T = \chi_h^{-1} \Phi_h^M \chi_h$ with $\chi_h = \Phi_{h/2}^E$, implying that the trapezoidal and midpoint rules are mutually conjugate. The change of coordinates, which transforms the numerical solution of one method to that of the other, is $\mathcal{O}(h)$ -close to the identity.

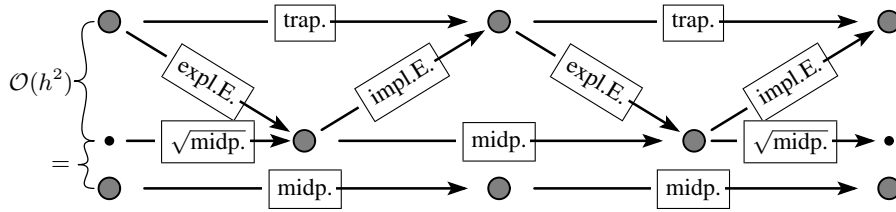


Fig. 8.1. Conjugacy of the trapezoidal rule and the implicit midpoint rule

In fact, we can do even better. If we let $\Phi_{h/2}$ be the square root of Φ_h^M (i.e., $\Phi_{h/2} \circ \Phi_{h/2} = \Phi_h^M$, see Lemma V.3.2), then we have (Fig. 8.1)

$$\Phi_h^T = (\Phi_{h/2}^E)^{-1} \circ \Phi_h^M \circ \Phi_{h/2}^E = (\Phi_{h/2}^E)^{-1} \circ \Phi_{h/2} \circ \Phi_{h/2} \circ \Phi_{h/2} \circ \Phi_{h/2}^{-1} \circ \Phi_{h/2}^E$$

so that the trapezoidal and the midpoint rules are conjugate via $\chi_h = \Phi_{h/2}^{-1} \circ \Phi_{h/2}^E$. Since $\Phi_{h/2}$ and $\Phi_{h/2}^E$ are both consistent with the same differential equation, the transformation χ_h is $\mathcal{O}(h^2)$ -close to the identity. This shows that for every numerical solution of the trapezoidal rule there exists a numerical solution of the midpoint rule which remains $\mathcal{O}(h^2)$ -close as long as it stays in a compact set. A single trajectory of the non-symplectic trapezoidal rule therefore behaves very much the same as a trajectory of the symplectic implicit midpoint rule.

A Study via B-Series. An investigation of Runge–Kutta methods, conjugate to a symplectic method, leads us to the following weaker requirement: we say that a numerical method Φ_h is *conjugate to a symplectic method Ψ_h up to order r* , if there exists a transformation $\chi_h(y) = y + \mathcal{O}(h)$ such that

$$\Phi_h(h) = (\chi_h^{-1} \circ \Psi_h \circ \chi_h)(y) + \mathcal{O}(h^{r+1}). \quad (8.2)$$

This implies that the error of such a method behaves as the superposition of the error of a symplectic method of order p with that of a non-symplectic method of order r .

In the following we assume that all methods considered as well as the conjugacy mapping χ_h can be represented as B-series

$$\Phi_h(y) = B(a, y), \quad \Psi_h(y) = B(b, y), \quad \chi_h(y) = B(c, y). \quad (8.3)$$

Using the composition formula (III.1.38) of B-series, condition (8.2) becomes

$$(ac)(\tau) = (cb)(\tau) \quad \text{for } |\tau| \leq r. \quad (8.4)$$

The following results are taken from the thesis of P. Leone (2000).

Theorem 8.2. *Let $\Phi_h(y) = B(a, y)$ represent a numerical method of order 2.*

a) It is always conjugate to a symplectic method up to order 3.

b) It is conjugate to a symplectic method up to order 4, if and only if

$$a(\bullet, \mathbf{V}) - 2a(\bullet, \mathbf{J}) = 0, \quad a(\mathbf{J}, \mathbf{J}) - 2a(\bullet, \mathbf{J}) = 0. \quad (8.5)$$

Here, we use the abbreviation $a(u, v) = a(u) \cdot a(v) - a(u \circ v) - a(v \circ u)$.

Proof. The condition (8.4) allows us to express $b(\tau)$ as a function of $a(u)$ for $|u| \leq |\tau|$ and of $c(v)$ for $|v| \leq |\tau| - 1$ (use the formulas of Example III.1.11). All we have to do is to check the symplecticity conditions $b(u, v) = 0$ for $|u| + |v| \leq r$ (see Theorem 7.6).

Since the method Φ_h is of order 2, we obtain $b(\bullet) = 1$ and $b(\mathbf{J}) = 1/2$. We arbitrarily fix $c(\bullet) = 0$, so that the symplecticity condition $b(\bullet, \mathbf{J}) = 0$ becomes $2c(\mathbf{J}) = a(\bullet, \mathbf{J})$. Defining $c(\mathbf{J})$ by this relation proves statement (a).

For order 4, the three symplecticity conditions $b(\bullet, \mathbf{V}) = b(\bullet, [[\bullet]]) = b(\mathbf{J}, \mathbf{J}) = 0$ have to be fulfilled. One of them can be satisfied by defining suitably $c(\mathbf{V}) + c([[\bullet]])$; the other two conditions are then equivalent to (8.5). \square

Theorem 8.3. *Let $\Phi_h(y) = B(a, y)$ represent a numerical method of order 4. It is conjugate to a symplectic method up to order 5, if and only if*

$$\begin{aligned} a(\bullet, \mathbf{V}) - 2a(\bullet, \mathbf{J}) &= 0, & a(\bullet, \mathbf{VV}) - 3a(\bullet, \mathbf{V}) + 3a(\bullet, \mathbf{J}) &= 0, \\ a(\mathbf{J}, \mathbf{V}) - a(\bullet, \mathbf{V}) - 2a(\mathbf{J}, \mathbf{J}) + 3a(\bullet, \mathbf{J}) &= 0. \end{aligned}$$

Proof. The idea of the proof is the same as in the preceding theorem. The verification is left as an exercise for the reader. \square

Example 8.4. A direct computation shows that for the Lobatto IIIB method with $s = 3$ we have $a(\mathbf{J}, \mathbf{V}) = 1/144$, and $a(u, v) = 0$ for all other pairs with $|u| + |v| = 5$. Theorem 8.3 therefore proves that this method is not conjugate to a symplectic method up to order 5.

For the Lobatto IIIA method with $s = 3$ we obtain $a(\mathbf{J}, \mathbf{V}) = -1/144$, $a(\mathbf{J}, [[\bullet]]) = -1/288$, and $a(u, v) = 0$ for the remaining pairs with $|u| + |v| = 5$. This time the conditions of Theorem 8.3 are fulfilled, so that the Lobatto IIIA method with $s = 3$ is conjugate to a symplectic method up to order 5 at least.

VI.8.2 Near Conservation of Quadratic First Integrals

We have already met in Sect. VI.4.1 a close relationship between symplecticity and the conservation of quadratic first integrals. The aim of this section is to show a similar connection between conjugate symplecticity and the near conservation of quadratic first integrals. This has first been observed and proved by Chartier, Faou & Murua (2005) using the algebra of rooted trees.

Let $Q(y) = y^T C y$ (with symmetric matrix C) be a quadratic first integral of $\dot{y} = f(y)$, and assume that $\Phi_h(y)$ is conjugate to a method $\Psi_h(y)$ that exactly conserves quadratic first integrals (e.g., symplectic Runge–Kutta methods). This means that $y_{n+1} = \Phi_h(y_n)$ satisfies

$$\chi_h(y_{n+1})^T C \chi_h(y_{n+1}) = \chi_h(y_n)^T C \chi_h(y_n),$$

and the expression $\tilde{Q}(y) = \chi_h(y)^T C \chi_h(y)$ is exactly conserved by the numerical solution of $\Phi_h(y)$. If $\chi_h(y) = B(c, y)$ is a B-series, this is of the form

$$\tilde{Q}(y) = \sum_{\tau, \vartheta \in T \cup \{\emptyset\}} h^{|\tau|+|\vartheta|} \beta(\tau, \vartheta) F(\tau)(y)^T C F(\vartheta)(y), \quad (8.6)$$

where $F(\emptyset)(y) = y$ and $|\emptyset| = 0$ for the empty tree, and $\beta(\emptyset, \emptyset) = 1$. We have the following criterion for conjugate symplecticity, where all formulas have to be interpreted in the sense of formal series.

Theorem 8.5. *Assume that a one-step method $\Phi_h(y) = B(a, y)$ leaves (8.6) invariant for all problems $\dot{y} = f(y)$ having $Q(y) = y^T C y$ as first integral.*

Then, it is conjugate to a symplectic integrator $\Psi_h(z)$, i.e., there exists a transformation $z = \chi_h(y) = B(c, y)$ such that $\Psi_h(z) = \chi_h \circ \Phi_h \circ \chi_h^{-1}(z)$, or equivalently, $\Psi_h(z) = B(c^{-1}ac, z)$ is symplectic.

Proof. The idea is to search for a B-series $B(c, y)$ such that the expression (8.6) becomes

$$\tilde{Q}(y) = B(c, y)^T C B(c, y).$$

The mapping $z = \chi_h(y) = B(c, y)$ then provides a change of variables such that the original first integral $Q(z) = z^T C z$ is invariant in the new variables. By Theorem 7.6 this then implies that Ψ_h is symplectic.

By Lemma 8.6 below, the expression (8.6) can be written as

$$\tilde{Q}(y) = y^T C \left(y + \sum_{\theta \in T} h^{|\theta|} \eta(\theta) F(\theta)(y) \right), \quad (8.7)$$

where $\eta(\theta) = 0$ for $|\theta| < r$, if the perturbation in (8.6) is of size $\mathcal{O}(h^r)$. Using the same lemma once more, we obtain

$$\begin{aligned} B(c, y)^T C B(c, y) &= y^T C \left(y + 2 \sum_{\theta \in T} \frac{h^{|\theta|}}{\sigma(\theta)} c(\theta) F(\theta)(y) \right) \\ &+ y^T C \left(\sum_{\theta \in T} \left(\frac{h^{|\theta|}}{\sigma(\theta)} \sum_{\tau, \vartheta \in T} \frac{\sigma(\theta) \kappa_{\tau, \vartheta}(\theta)}{\sigma(\tau) \sigma(\vartheta)} c(\tau) c(\vartheta) F(\theta)(y) \right) \right). \end{aligned} \quad (8.8)$$

A comparison of the coefficients in (8.7) and (8.8) uniquely defines $c(\theta)$ in a recursive manner. We have $c(\theta) = 0$ for $|\theta| < r$, so that the transformation $z = B(c, y)$ is $\mathcal{O}(h^r)$ close to the identity. \square

The previous proof is based on the following result.

Lemma 8.6. *Let $Q(y) = y^T C y$ (with symmetric matrix C) be a first integral of $\dot{y} = f(y)$. Then, for every pair of trees $\tau, \vartheta \in T$, we have*

$$F(\tau)(y)^T C F(\vartheta)(y) = y^T C \left(\sum_{\theta \in T} \kappa_{\tau, \vartheta}(\theta) F(\theta)(y) \right).$$

This sum is finite and only over trees satisfying $|\theta| = |\tau| + |\vartheta|$.

Proof. By definition of a first integral we have $y^T C f(y) = 0$ for all y . Differentiation with respect to y gives

$$f(y)^T C k + y^T C f'(y)k = 0 \quad \text{for all } k. \quad (8.9)$$

Putting $k = F(\vartheta)(y)$, this proves the statement for $\tau = \bullet$.

Differentiating once more yields

$$(f'(y)\ell)^T C k + \ell^T C f'(y)k + y^T C f''(y)(k, \ell) = 0.$$

Putting $\ell = f(y)$ and using (8.9), we get the statement for $\tau = \text{f}$. With $\ell = F(\tau_1)(y)$ we obtain the statement for $\tau = [\tau_1]$ provided that it is already proved for τ_1 . We need a further differentiation to get a similar statement for $\tau = [\tau_1, \tau_2]$, etc. The proof concludes by induction on the order of τ . \square

Partitioned Methods. This criterion for conjugate symplecticity can be extended to partitioned P-series methods. For partitioned problems

$$\dot{p} = f_1(p, q), \quad \dot{q} = f_2(p, q) \quad (8.10)$$

we consider first integrals of the form $L(p, q) = p^T E q$, where E is an arbitrary constant matrix. If $\Phi_h(p, q)$ is conjugate to a method that exactly conserves $L(p, q)$, then it will conserve a modified first integral of the form

$$\tilde{L}(p, q) = \sum_{\tau \in TP_p \cup \{\emptyset_p\}, \vartheta \in TP_q \cup \{\emptyset_q\}} h^{|\tau|+|\vartheta|} \beta(\tau, \vartheta) F(\tau)(p, q)^T E F(\vartheta)(p, q), \quad (8.11)$$

where $\beta(\emptyset_p, \emptyset_q) = 1$, $F(\emptyset_p)(p, q) = p$, $F(\emptyset_q)(p, q) = q$. We first extend Lemma 8.6 to the new situation.

Lemma 8.7. *Let $L(p, q) = p^T E q$ be a first integral of (8.10). Then, for every pair of trees $\tau \in TP_p, \vartheta \in TP_q$, we have*

$$\begin{aligned} F(\tau)(p, q)^T E F(\vartheta)(p, q) &= p^T E \left(\sum_{\theta \in TP_q} \kappa_{\tau, \vartheta}(\theta) F(\theta)(p, q) \right) \\ &+ \left(\sum_{\theta \in TP_p} \kappa_{\tau, \vartheta}(\theta) F(\theta)(p, q) \right)^T E q. \end{aligned} \quad (8.12)$$

These sums are finite and only over trees satisfying $|\theta| = |\tau| + |\vartheta|$.

Proof. Since $L(p, q) = p^T E q$ is a first integral of the differential equation, we have $f_1(p, q)^T E q + p^T E f_2(p, q) = 0$ for all p and q . As in the proof of Lemma 8.6 the statement follows from differentiation of this relation. \square

Theorem 8.8. *Assume that a partitioned one-step method $\Phi_h(p, q) = P(a, (p, q))$ leaves (8.11) invariant for all problems (8.10) having $L(p, q) = p^T E q$ as first integral.*

Then it is conjugate to a symplectic integrator $\Psi_h(u, v)$, i.e., there is a transformation $(u, v) = \chi_h(p, q) = P(c, (p, q))$ such that $\Psi_h(u, v) = \chi_h \circ \Phi_h \circ \chi_h^{-1}(u, v)$, or equivalently, $\Psi_h(u, v) = P(c^{-1} a c, (u, v))$ is symplectic.

Proof. We search for a P-series $P(c, (p, q)) = (P_p(c, (p, q)), P_q(c, (p, q)))^T$ such that the expression (8.11) can be written as

$$\tilde{L}(p, q) = P_p(c, (p, q))^T E P_q(c, (p, q)).$$

As in the proof of Theorem 8.5 the mapping $(u, v) = \chi_h(p, q) = P(c, (p, q))$ then provides the searched change of variables.

Using Lemma 8.7 the expression (8.11) becomes

$$\tilde{L}(p, q) = p^T E \left(q + \sum_{\theta \in TP_q} h^{|\theta|} \eta(\theta) F(\theta)(p, q) \right) + \left(\sum_{\theta \in TP_p} h^{|\theta|} \eta(\theta) F(\theta)(p, q) \right)^T E q.$$

Also $P_p(c, (p, q))^T E P_q(c, (p, q))$ can be written in such a form, and a comparison of the coefficients yields the coefficients $c(\tau)$ of the P-series $P(c, (p, q))$ in a recursive manner. We again have that $P(c, (p, q))$ is $\mathcal{O}(h^r)$ close to the identity, if the perturbation in (8.11) is of size $\mathcal{O}(h^r)$. \square

The statement of Theorem 8.8 remains true in the class of second order differential equations $\ddot{q} = f_1(q)$, i.e., $\dot{p} = f_1(p)$, $\dot{q} = p$.

VI.9 Volume Preservation

The flow φ_t of a Hamiltonian system preserves volume in phase space: for every bounded open set $\Omega \subset \mathbb{R}^{2d}$ and for every t for which $\varphi_t(y)$ exists for all $y \in \Omega$,

$$\text{vol}(\varphi_t(\Omega)) = \text{vol}(\Omega),$$

where $\text{vol}(\Omega) = \int_{\Omega} dy$. This identity is often referred to as *Liouville's theorem*. It is a consequence of the transformation formula for integrals and the fact that

$$\det \frac{\partial \varphi_t(y)}{\partial y} = 1 \quad \text{for all } t \text{ and } y, \quad (9.1)$$

which follows directly from the symplecticity and $\varphi_0 = \text{id}$. The same argument shows that every symplectic transformation, and in particular every symplectic integrator applied to a Hamiltonian system, preserves volume in phase space.

More generally than for Hamiltonian systems, volume is preserved by the flow of differential equations with a divergence-free vector field:

Lemma 9.1. *The flow of a differential equation $\dot{y} = f(y)$ in \mathbb{R}^n is volume-preserving if and only if $\operatorname{div} f(y) = 0$ for all y .*

Proof. The derivative $Y(t) = \frac{\partial \varphi_t}{\partial y}(y_0)$ is the solution of the variational equation

$$\dot{Y} = A(t)Y, \quad Y(0) = I,$$

with the Jacobian matrix $A(t) = f'(y(t))$ at $y(t) = \varphi_t(y_0)$. From the proof of Lemma IV.3.1 we obtain the *Abel–Liouville–Jacobi–Ostrogradskii identity*

$$\frac{d}{dt} \det Y = \operatorname{trace} A(t) \cdot \det Y. \quad (9.2)$$

Note that here $\operatorname{trace} A(t) = \operatorname{div} f(y(t))$. Hence, $\det Y(t) = 1$ for all t if and only if $\operatorname{div} f(y(t)) = 0$ for all t . Since this is valid for all choices of initial values y_0 , the result follows. \square

Example 9.2 (ABC Flow). This flow, named after the three independent authors Arnold, Beltrami and Childress, is given by the equations

$$\begin{aligned} \dot{x} &= A \sin z + C \cos y \\ \dot{y} &= B \sin x + A \cos z \\ \dot{z} &= C \sin y + B \cos x \end{aligned} \quad (9.3)$$

and has all diagonal elements of f' identically zero. It is therefore volume preserving. In Arnold (1966, p. 347) it appeared in a footnote as an example of a flow with $\operatorname{rot} f$ parallel to f , thus violating Arnold's condition for the existence of invariant tori (Arnold 1966, p. 346). It was therefore expected to possess interesting chaotic properties and has since then been the object of many investigations showing their non-integrability (see e.g., Ziglin (1996)). We illustrate in Fig. 9.1 the action of this flow by transforming, in a volume preserving manner, a ball in \mathbb{R}^3 . We see that, very soon, the set is strongly squeezed in one direction and dilated in two others. The solutions thus depend in a very sensitive way on the initial values.

Volume-Preserving Numerical Integrators. The question arises as to whether volume-preserving integrators can be constructed for every differential equation with volume-preserving flow. Already for linear problems, Lemma IV.3.2 shows that no standard method can be volume-preserving for dimension $n \geq 3$. Nevertheless, positive answers were found by Qin & Zhu (1993), Shang (1994a, 1994b), Feng & Shang (1995) and Quispel (1995). In the following we present the approach of Feng & Shang (1995). The key is the following result which generalizes and reinterprets a construction of H. Weyl (1940) for $n = 3$.

Theorem 9.3 (Feng & Shang 1995). *Every divergence-free vector field $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ can be written as the sum of $n - 1$ vector fields*

$$f = f_{1,2} + f_{2,3} + \dots + f_{n-1,n}$$

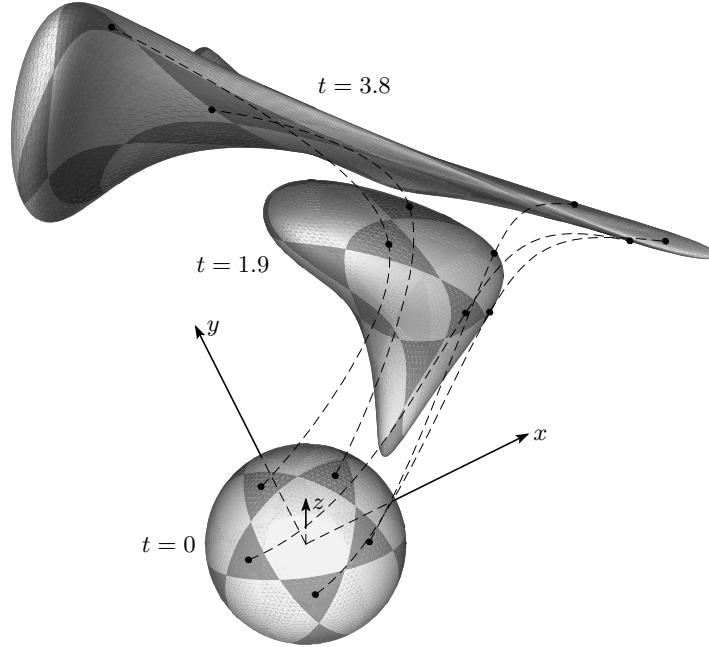


Fig. 9.1. Volume preserving deformation of the ball of radius 1, centred at the origin, by the ABC flow; $A = 1/2$, $B = C = 1$

where each $f_{k,k+1}$ is Hamiltonian in the variables (y_k, y_{k+1}) : there exist functions $H_{k,k+1} : \mathbb{R}^n \rightarrow \mathbb{R}$ such that

$$f_{k,k+1} = (0, \dots, 0, -\frac{\partial H_{k,k+1}}{\partial y_{k+1}}, \frac{\partial H_{k,k+1}}{\partial y_k}, 0, \dots, 0)^T.$$

Proof. In terms of the components of $f = (f_1, \dots, f_n)^T$, the functions $H_{k,k+1}$ must satisfy the equations

$$\begin{aligned} f_1 &= -\frac{\partial H_{1,2}}{\partial y_2}, & f_2 &= \frac{\partial H_{1,2}}{\partial y_1} - \frac{\partial H_{2,3}}{\partial y_3}, \dots, \\ f_{n-1} &= \frac{\partial H_{n-2,n-1}}{\partial y_{n-2}} - \frac{\partial H_{n-1,n}}{\partial y_n}, & f_n &= \frac{\partial H_{n-1,n}}{\partial y_{n-1}}. \end{aligned}$$

We thus set

$$H_{1,2} = -\int_0^{y_2} f_1 dy_2$$

and for $k = 2, \dots, n-2$

$$H_{k,k+1} = \int_0^{y_{k+1}} \left(\frac{\partial H_{k-1,k}}{\partial y_{k-1}} - f_k \right) dy_{k+1}.$$

It remains to construct $H_{n-1,n}$ from the last two equations. We see by induction that for $k \leq n-2$,

$$\frac{\partial^2 H_{k,k+1}}{\partial y_k \partial y_{k+1}} = -\left(\frac{\partial f_1}{\partial y_1} + \dots + \frac{\partial f_k}{\partial y_k}\right),$$

and hence the integrability condition for $H_{n-1,n}$,

$$\frac{\partial}{\partial y_{n-1}} \left(\frac{\partial H_{n-2,n-1}}{\partial y_{n-2}} - f_{n-1} \right) = \frac{\partial f_n}{\partial y_n},$$

reduces to the condition $\operatorname{div} f = 0$, which is satisfied by assumption. $H_{n-1,n}$ can thus be constructed as

$$H_{n-1,n} = \int_0^{y_n} \left(\frac{\partial H_{n-2,n-1}}{\partial y_{n-2}} - f_{n-1} \right) dy_n + \int_0^{y_{n-1}} f_n|_{y_n=0} dy_{n-1},$$

which completes the proof. \square

The above construction also shows that

$$f_{k,k+1} = (0, \dots, 0, f_k + g_k, -g_{k+1}, 0, \dots, 0)$$

with

$$g_{k+1} = \int_0^{y_{k+1}} \left(\frac{\partial f_1}{\partial y_1} + \dots + \frac{\partial f_k}{\partial y_k} \right) dy_{k+1}$$

for $1 \leq k \leq n-2$, and $g_1 = 0$ and $g_n = -f_n$.

With the decomposition of Lemma 9.3 at hand, a volume-preserving algorithm is obtained by applying a splitting method with symplectic substeps. For example, as proposed by Feng & Shang (1995), a second-order volume-preserving method is obtained by Strang splitting with symplectic Euler substeps:

$$\varphi_h \approx \Phi_h = \Phi_{h/2}^{[1,2]*} \circ \dots \circ \Phi_{h/2}^{[n-1,n]*} \circ \Phi_{h/2}^{[n-1,n]} \circ \dots \circ \Phi_{h/2}^{[1,2]}$$

where $\Phi_{h/2}^{[k,k+1]}$ is a symplectic Euler step of length $h/2$ applied to the system with right-hand side $f_{k,k+1}$, and $*$ denotes the adjoint method. In this method, one step $\hat{y} = \Phi_h(y)$ is computed component-wise, in a Gauss-Seidel-like manner, as

$$\begin{aligned} \bar{y}_1 &= y_1 + \frac{h}{2} f_1(\bar{y}_1, y_2, \dots, y_n) \\ \bar{y}_k &= y_k + \frac{h}{2} f_k(\bar{y}_1, \dots, \bar{y}_k, y_{k+1}, \dots, y_n) + \frac{h}{2} g_k|_{\bar{y}_k} \quad \text{for } k = 2, \dots, n-1 \\ \bar{y}_n &= y_n + \frac{h}{2} f_n(\bar{y}_1, \dots, \bar{y}_{n-1}, y_n) \end{aligned} \tag{9.4}$$

with $g_k|_{\bar{y}_k} = g_k(\bar{y}_1, \dots, \bar{y}_k, y_{k+1}, \dots, y_n) - g_k(\bar{y}_1, \dots, \bar{y}_{k-1}, y_k, \dots, y_n)$, and

$$\begin{aligned}
\widehat{y}_n &= \overline{y}_n + \frac{h}{2} f_n(\overline{y}_1, \dots, \widehat{y}_n) \\
\widehat{y}_k &= \overline{y}_k + \frac{h}{2} f_k(\overline{y}_1, \dots, \overline{y}_k, \widehat{y}_{k+1}, \dots, \widehat{y}_n) - \frac{h}{2} \overline{g}_k \Big|_{\overline{y}_k}^{\widehat{y}_k} \quad \text{for } k = n-1, \dots, 2 \\
\widehat{y}_1 &= \overline{y}_1 + \frac{h}{2} f_1(\overline{y}_1, \widehat{y}_2, \dots, \widehat{y}_n)
\end{aligned} \tag{9.5}$$

with $\overline{g}_k \Big|_{\overline{y}_k}^{\widehat{y}_k} = g_k(\overline{y}_1, \dots, \overline{y}_{k-1}, \widehat{y}_k, \dots, \widehat{y}_n) - g_k(\overline{y}_1, \dots, \overline{y}_k, \widehat{y}_{k+1}, \dots, \widehat{y}_n)$. The method is one-dimensionally implicit in general, but becomes explicit in the particular case where $\partial f_k / \partial y_k = 0$ for all k .

Separable Partitioned Systems. For problems of the form

$$\dot{y} = f(z), \quad \dot{z} = g(y) \tag{9.6}$$

with $y \in \mathbb{R}^m$, $z \in \mathbb{R}^n$, the scheme (9.4) becomes the symplectic Euler method, (9.5) its adjoint, and its composition the Lobatto IIIA - IIIB extension of the Störmer–Verlet method. Since symplectic explicit partitioned Runge–Kutta methods are compositions of symplectic Euler steps (Theorem VI.4.7), this observation proves that such methods are volume-preserving for systems (9.6). This fact was obtained by Suris (1996) by a direct calculation, without interpreting the methods as composition methods. The question arises as to whether more symplectic partitioned Runge–Kutta methods are volume-preserving for systems (9.6).

Theorem 9.4. *Every symplectic Runge–Kutta method with at most two stages is volume-preserving for systems (9.6) of arbitrary dimension.*

Proof. (a) The idea is to consider the Hamiltonian system with

$$H(u, v, y, z) = u^T f(z) + v^T g(y),$$

where (u, v) are the conjugate variables to (y, z) . This system is of the form

$$\begin{aligned}
\dot{y} &= f(z) & \dot{u} &= -g'(y)^T v \\
\dot{z} &= g(y) & \dot{v} &= -f'(z)^T u.
\end{aligned} \tag{9.7}$$

Applying the Runge–Kutta method to this augmented system does not change the numerical solution for (y, z) . For symplectic methods the matrix

$$\left(\frac{\partial(y_1, z_1, u_1, v_1)}{\partial(y_0, z_0, u_0, v_0)} \right) = M = \begin{pmatrix} R & 0 \\ S & T \end{pmatrix} \tag{9.8}$$

satisfies $M^T J M = J$ which implies $R T^T = I$. Below we shall show that $\det T = \det R$. This yields $\det R = 1$ which implies that the method is volume preserving.

(b) *One-stage methods.* The only symplectic one-stage method is the implicit midpoint rule for which R and T are computed as

$$\left(I - \frac{h}{2} E_1\right) R = I + \frac{h}{2} E_1 \quad (9.9)$$

$$\left(I + \frac{h}{2} E_1^T\right) T = I - \frac{h}{2} E_1^T, \quad (9.10)$$

where E_1 is the Jacobian of the system (9.6) evaluated at the internal stage value. Since

$$E_1 = \begin{pmatrix} 0 & f'(z_{1/2}) \\ g'(y_{1/2}) & 0 \end{pmatrix},$$

a similarity transformation with the matrix $D = \text{diag}(I, -I)$ takes E_1 to $-E_1$. Hence, the transformed matrix satisfies

$$\left(I - \frac{h}{2} E_1^T\right) (D^{-1} T D) = I + \frac{h}{2} E_1^T.$$

A comparison with (9.9) and the use of $\det X^T = \det X$ proves $\det R = \det T$ for the midpoint rule.

(c) *Two-stage methods.* Applying a two-stage implicit Runge–Kutta method to (9.7) yields

$$\begin{pmatrix} I - ha_{11}E_1 & -ha_{12}E_2 \\ -ha_{21}E_1 & I - ha_{22}E_2 \end{pmatrix} \begin{pmatrix} R_1 \\ R_2 \end{pmatrix} = \begin{pmatrix} I \\ I \end{pmatrix},$$

where R_i is the derivative of the (y, z) components of the i th stage with respect to (y_0, z_0) , and E_i is the Jacobian of the system (9.6) evaluated at the i th internal stage value. From the solution of this system the derivative R of (9.8) is obtained as

$$R = I + (b_1 E_1, b_2 E_2) \begin{pmatrix} I - ha_{11}E_1 & -ha_{12}E_2 \\ -ha_{21}E_1 & I - ha_{22}E_2 \end{pmatrix}^{-1} \begin{pmatrix} I \\ I \end{pmatrix}.$$

With the determinant identity

$$\det(U) \det(X - WU^{-1}V) = \det \begin{pmatrix} U & V \\ W & X \end{pmatrix} = \det(X) \det(U - VX^{-1}W),$$

which is seen by Gaussian elimination, this yields

$$\det R = \frac{\det(I \otimes I - h((A - \mathbb{1}b^T) \otimes I) E)}{\det(I \otimes I - h(A \otimes I) E)},$$

where A and b collect the Runge–Kutta coefficients, and $E = \text{blockdiag}(E_1, E_2)$. For $D^{-1}TD$ we get the same formula with E replaced by E^T . If A is an arbitrary 2×2 matrix, it follows from block Gaussian elimination that

$$\det(I \otimes I - h(A \otimes I) E) = \det(I \otimes I - h(A \otimes I) E^T), \quad (9.11)$$

which then proves $\det R = \det T$. Notice that the identity (9.11) is no longer true in general if A is of dimension larger than two. \square

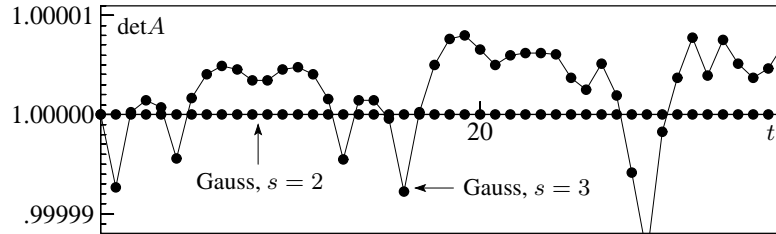


Fig. 9.2. Volume preservation of Gauss methods applied to (9.12) with $h = 0.8$

We are curious to see whether Theorem 9.4 remains valid for symplectic Runge–Kutta methods with more than two stages. For this we apply the Gauss methods with $s = 2$ and $s = 3$ to the problem

$$\dot{x} = \sin z, \quad \dot{y} = \cos z, \quad \dot{z} = \sin y + \cos x \quad (9.12)$$

with initial value $(0, 0, 0)$. We show in Fig. 9.2 the determinant of the derivative of the numerical flow as a function of time. Only the two-stage method is volume-preserving for this problem which is in agreement with Theorem 9.4.

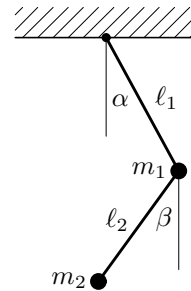
VI.10 Exercises

- Let α and β be the generalized coordinates of the double pendulum, whose kinetic and potential energies are

$$T = \frac{m_1}{2}(\dot{x}_1^2 + \dot{y}_1^2) + \frac{m_2}{2}(\dot{x}_2^2 + \dot{y}_2^2)$$

$$U = m_1 g y_1 + m_2 g y_2.$$

Determine the generalized momenta of the corresponding Hamiltonian system.



- A non-autonomous Hamiltonian system is given by a time-dependent Hamiltonian function $H(p, q, t)$ and the differential equations

$$\dot{p} = -H_q(p, q, t), \quad \dot{q} = H_p(p, q, t).$$

Verify that these equations together with $\dot{e} = -H_t(p, q, t)$ and $\dot{t} = 1$ are the canonical equations for the extended Hamiltonian $\tilde{H}(\tilde{p}, \tilde{q}) = H(p, q, t) + e$ with $\tilde{p} = (p, e)$ and $\tilde{q} = (q, t)$.

- Prove that a linear transformation $A : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is symplectic, if and only if $\det A = 1$.
- Consider the transformation $(r, \varphi) \mapsto (p, q)$, defined by

$$p = \psi(r) \cos \varphi, \quad q = \psi(r) \sin \varphi.$$

For which function $\psi(r)$ is it a symplectic transformation?

5. Prove that the definition (2.4) of $\Omega(M)$ does not depend on the parametrization φ , i.e., the parametrization $\psi = \varphi \circ \alpha$, where α is a diffeomorphism between suitable domains of \mathbb{R}^2 , leads to the same result.
6. On the set $U = \{(p, q) ; p^2 + q^2 > 0\}$ consider the differential equation

$$\begin{pmatrix} \dot{p} \\ \dot{q} \end{pmatrix} = \frac{1}{p^2 + q^2} \begin{pmatrix} p \\ q \end{pmatrix}. \quad (10.1)$$

Prove that

- a) its flow is symplectic everywhere on U ;
- b) on every simply-connected subset of U the vector field (10.1) is Hamiltonian (with $H(p, q) = -\text{Im} \log(p + iq) + \text{Const}$);
- c) it is not possible to find a differentiable function $H : U \rightarrow \mathbb{R}$ such that (10.1) is equal to $J^{-1}\nabla H(p, q)$ for all $(p, q) \in U$.

Remark. The vector field (10.1) is locally (but not globally) Hamiltonian.

7. (Burnton & Scherer 1998). Prove that all members of the one-parameter family of Nyström methods of order $2s$, constructed in Exercise III.9, are symplectic and symmetric.
8. Prove that the statement of Lemma 4.1 remains true for methods that are formally defined by a B-series, $\Phi_h(y) = B(a, y)$.
9. Compute the generating function $S^1(P, q, h)$ of a symplectic Nyström method applied to $\ddot{q} = U(q)$.
10. Find the Hamilton–Jacobi equation (cf. Theorem 5.7) for the generating function $S^2(p, Q)$ of Lemma 5.3.
11. (*Jacobi’s method for exact integration*). Suppose we have a solution $S(q, Q, t, \alpha)$ of the Hamilton–Jacobi equation (5.16), depending on d parameters $\alpha_1, \dots, \alpha_d$ such that the matrix $\left(\frac{\partial^2 S}{\partial \alpha_i \partial Q_j}\right)$ is invertible. Since this matrix is the Jacobian of the system

$$\frac{\partial S}{\partial \alpha_i} = 0 \quad i = 1, \dots, d, \quad (10.2)$$

this system determines a solution path Q_1, \dots, Q_d which is locally unique. In possession of an additional parameter (and, including the partial derivatives with respect to t , an additional row and column in the Hessian matrix condition), we can also determine $Q_j(t)$ as function of t . Apply this method to the Kepler problem (I.2.2) in polar coordinates, where, with the generalized momenta $p_r = \dot{r}$, $p_\varphi = r^2 \dot{\varphi}$, the Hamiltonian becomes

$$H = \frac{1}{2} \left(p_r^2 + \frac{p_\varphi^2}{r^2} \right) - \frac{M}{r}$$

and the Hamilton–Jacobi differential equation (5.16) is

$$S_t + \frac{1}{2}(S_r)^2 + \frac{1}{2r^2}(S_\varphi)^2 - \frac{M}{r} = 0.$$

Solve this equation by the ansatz $S(t, r, \varphi) = \theta_1(t) + \theta_2(r) + \theta_3(\varphi)$ (separation of variables).

Result. One obtains

$$S = \int \sqrt{2\alpha_1 r^2 + 2Mr - \alpha_2^2} \frac{dr}{r} + \alpha_2 \varphi - \alpha_1 t.$$

Putting, e.g., $\partial S / \partial \alpha_2 = 0$, we obtain $\varphi = \arcsin \frac{Mr - \alpha_2^2}{\sqrt{M^2 + 2\alpha_1 \alpha_2^2} r}$ by evaluating an elementary integral. This, when resolved for r , leads to the elliptic movement of Kepler (Sect. I.2.2). This method turned out to be most effective for the exact integration of difficult problems. With the same ideas, just more complicated in the computations, Jacobi solves in “lectures” 24 through 30 of (Jacobi 1842) the Kepler motion in \mathbb{R}^3 , the geodesics of ellipsoids (his greatest triumph), the motion with two centres of gravity, and proves a theorem of Abel.

12. (*Chan’s Lobatto IIIS methods.*) Show that there exists a one-parameter family of symplectic, symmetric (and A -stable) Runge–Kutta methods of order $2s - 2$ based on Lobatto quadrature (Chan 1990). A special case of these methods can be obtained by taking the arithmetic mean of the Lobatto IIIA and Lobatto IIIB method coefficients (Sun 2000).

Hint. Use the W -transformation (see Hairer & Wanner (1996), p. 77) by putting $X_{s,s-1} = -X_{s-1,s}$ an arbitrary constant.

13. For a Hamiltonian system with associated Lagrangian $L(q, \dot{q}) = \frac{1}{2} \dot{q}^T M \dot{q} - U(q)$, show that every first integral $I(p, q) = p^T a(q)$ resulting from Noether’s Theorem has a linear $a(q) = Aq + c$ with skew-symmetric MA .

Hint. (a) It is sufficient to consider the case $M = I$.

(b) Show that $a'(q)$ is skew-symmetric.

(c) Let $a_{ij}(q) = \frac{\partial a_i}{\partial q_j}(q)$. Using the symmetry of the Hessian of each component $a_i(q)$, show that $a_{ij}(q)$ does not depend on q_i, q_j , and is at most linear in the remaining components q_k . With the skew-symmetry of $a'(q)$, conclude that $a'(q) = \text{Const.}$

14. Consider the unconstrained *optimal control problem*

$$\begin{aligned} C(q(T)) &\rightarrow \min \\ \dot{q}(t) &= f(q(t), u(t)), \quad q(0) = q_0 \end{aligned} \quad (10.3)$$

on the interval $[0, T]$, where the control function is assumed to be continuous.

Prove that first-order necessary optimality conditions can be written as

$$\begin{aligned} \dot{q}(t) &= \nabla_p H(p(t), q(t), u(t)), & q(0) &= q_0 \\ \dot{p}(t) &= -\nabla_q H(p(t), q(t), u(t)), & p(T) &= \nabla_q C(q(T)) \\ 0 &= \nabla_u H(p(t), q(t), u(t)), \end{aligned} \quad (10.4)$$

where the Hamiltonian is given by

$$H(p, q, u) = p^T f(q, u)$$

(we assume that the Hessian $\nabla_u^2 H(p, q, u)$ is invertible, so that the third relation of (10.4) defines u as a function of (p, q)).

Hint. Consider a slightly perturbed control function $u(t) + \varepsilon \delta u(t)$, and let $q(t) + \varepsilon \delta q(t) + \mathcal{O}(\varepsilon^2)$ be the corresponding solution of the differential equation in (10.3). With the function $p(t)$ of (10.4) we then have

$$C'(q(T)) \delta q(T) = \int_0^T \frac{d}{dt} \left(p(t)^T \delta q(t) \right) dt = \int_0^T p(t)^T f_u(\dots) \delta u(t) dt.$$

The algebraic relation of (10.4) then follows from the fundamental lemma of variational calculus.

15. A Runge–Kutta discretization of the problem (10.3) is

$$\begin{aligned} C(q_N) &\rightarrow \min \\ q_{n+1} &= q_n + h \sum_{i=1}^s b_i f(Q_{ni}, U_{ni}) \\ Q_{ni} &= q_n + h \sum_{j=1}^s a_{ij} f(Q_{nj}, U_{nj}) \end{aligned} \quad (10.5)$$

with $n = 0, \dots, N-1$ and $h = T/N$. We assume $b_i \neq 0$ for all i . Introducing suitable Lagrange multipliers for the constrained minimization problem (10.5), prove that there exist p_n, P_{ni} such that the optimal solution of (10.5) satisfies (Hager 2000)

$$\begin{aligned} q_{n+1} &= q_n + h \sum_{i=1}^s b_i \nabla_p H(P_{ni}, Q_{ni}, U_{ni}) \\ Q_{ni} &= q_n + h \sum_{j=1}^s a_{ij} \nabla_p H(P_{nj}, Q_{nj}, U_{nj}) \\ p_{n+1} &= p_n - h \sum_{i=1}^s \hat{b}_i \nabla_q H(P_{ni}, Q_{ni}, U_{ni}) \\ P_{ni} &= p_n - h \sum_{j=1}^s \hat{a}_{ij} \nabla_q H(P_{nj}, Q_{nj}, U_{nj}) \\ 0 &= \nabla_u H(P_{ni}, Q_{ni}, U_{ni}) \end{aligned} \quad (10.6)$$

with $p_N = \nabla_q C(q_N)$ and given initial value q_0 , where the coefficients \hat{b}_i and \hat{a}_{ij} are determined by

$$\hat{b}_i = b_i, \quad b_i \hat{a}_{ij} + \hat{b}_j a_{ji} = b_i \hat{b}_j. \quad (10.7)$$

Consequently, (10.6) can be considered as a symplectic discretization of (10.4); see Bonnans & Laurent-Varin (2006).

16. (Hager 2000). For an explicit s -stage Runge–Kutta method of order $p = s$ and $b_i \neq 0$, consider the partitioned Runge–Kutta method with additional coefficients \hat{b}_i and \hat{a}_{ij} defined by (10.7). Prove the following:
- For $p = s = 3$, the partitioned method is of order 3 if and only if $c_3 = 1$.
 - For $p = s = 4$, the partitioned method is of order 4 without any restriction.

Chapter VII.

Non-Canonical Hamiltonian Systems

We discuss theoretical properties and the structure-preserving numerical treatment of Hamiltonian systems on manifolds and of the closely related class of Poisson systems. We present numerical integrators for problems from classical and quantum mechanics.

VII.1 Constrained Mechanical Systems

Constrained mechanical systems form an important class of differential equations on manifolds. Their numerical treatment has been extensively investigated in the context of *differential-algebraic equations* and is documented in monographs like that of Brenan, Campbell & Petzold (1996), Eich-Soellner & Führer (1998), Hairer, Lubich & Roche (1989), and Chap. VII of Hairer & Wanner (1996). We concentrate here on the symmetry and/or symplecticity of such numerical integrators.

VII.1.1 Introduction and Examples

Consider a mechanical system described by position coordinates q_1, \dots, q_d , and suppose that the motion is constrained to satisfy $g(q) = 0$ where $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$ with $m < d$. Let $T(q, \dot{q}) = \frac{1}{2} \dot{q}^T M(q) \dot{q}$ be the kinetic energy of the system and $U(q)$ its potential energy, and put

$$L(q, \dot{q}) = T(q, \dot{q}) - U(q) - g(q)^T \lambda, \quad (1.1)$$

where $\lambda = (\lambda_1, \dots, \lambda_m)^T$ consists of Lagrange multipliers. The Euler–Lagrange equation of the variational problem for $\int_0^t L(q, \dot{q}) dt$ is then given by

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}} \right) - \frac{\partial L}{\partial q} = 0.$$

Written as a first order differential equation we get

$$\begin{aligned} \dot{q} &= v \\ M(q) \dot{v} &= f(q, v) - G(q)^T \lambda \\ 0 &= g(q), \end{aligned} \quad (1.2)$$

where $f(q, v) = -\frac{\partial}{\partial q} (M(q)v)v + \nabla_q T(q, v) - \nabla_q U(q)$ and $G(q) = \frac{\partial g}{\partial q}(q)$.

Example 1.1 (Spherical Pendulum). We denote by q_1, q_2, q_3 the Cartesian coordinates of a point with mass m that is connected with a massless rod of length ℓ to the origin. The kinetic and potential energies are $T = \frac{m}{2}(\dot{q}_1^2 + \dot{q}_2^2 + \dot{q}_3^2)$ and $U = mgq_3$, respectively, and the constraint is the fixed length of the rod. We thus get the system

$$\begin{aligned}\dot{q}_1 &= v_1 & m\dot{v}_1 &= -2q_1\lambda \\ \dot{q}_2 &= v_2 & m\dot{v}_2 &= -2q_2\lambda \\ \dot{q}_3 &= v_3 & m\dot{v}_3 &= -mg - 2q_3\lambda \\ 0 &= q_1^2 + q_2^2 + q_3^2 - \ell^2.\end{aligned}\tag{1.3}$$

The physical meaning of λ is the tension in the rod which maintains the constant distance of the mass point from the origin.

Existence and Uniqueness of the Solution. A standard approach for studying the existence of solutions of differential-algebraic equations is to differentiate the constraints until an ordinary differential equation is obtained. Differentiating the constraint in (1.2) twice with respect to time yields

$$0 = G(q)v \quad \text{and} \quad 0 = g''(q)(v, v) + G(q)\dot{v}.\tag{1.4}$$

The equation for \dot{v} in (1.2) together with the second relation of (1.4) constitute a linear system for \dot{v} and λ ,

$$\begin{pmatrix} M(q) & G(q)^T \\ G(q) & 0 \end{pmatrix} \begin{pmatrix} \dot{v} \\ \lambda \end{pmatrix} = \begin{pmatrix} f(q, v) \\ -g''(q)(v, v) \end{pmatrix}.\tag{1.5}$$

Throughout this chapter we require the matrix appearing in (1.5) to be invertible for q close to the solution we are looking for. This then allows us to express \dot{v} and λ as functions of (q, v) . Notice that the matrix in (1.5) is invertible when $G(q)$ has full rank and $M(q)$ is invertible on $\ker G(q) = \{h \mid G(q)h = 0\}$.

We are now able to discuss the existence of a solution of (1.2). First of all, observe that the initial values q_0, v_0, λ_0 cannot be arbitrarily chosen. They have to satisfy the first relation of (1.4) and $\lambda_0 = \lambda(q_0, v_0)$, where $\lambda(q, v)$ is obtained from (1.5). In the case that q_0, v_0, λ_0 satisfy these conditions, we call them *consistent initial values*. Furthermore, every solution of (1.2) has to satisfy

$$\dot{q} = v, \quad \dot{v} = \dot{v}(q, v),\tag{1.6}$$

where $\dot{v}(q, v)$ is the function obtained from (1.5). It is known from standard theory of ordinary differential equations that (1.6) has locally a unique solution. This solution $(q(t), v(t))$ together with $\lambda(t) := \lambda(q(t), v(t))$ satisfies (1.5) by construction, and hence also the two differential equations of (1.2). Integrating the second relation of (1.4) twice and using the fact that the integration constants vanish for consistent initial values, proves also the remaining relation $0 = g(q)$ for this solution.

Formulation as a Differential Equation on a Manifold. We denote by

$$\mathcal{Q} = \{q; g(q) = 0\} \quad (1.7)$$

the *configuration manifold*, on which the positions q are constrained to lie. The tangent space at $q \in \mathcal{Q}$ is $T_q\mathcal{Q} = \{v; G(q)v = 0\}$. The equations (1.6) define thus a differential equation on the manifold

$$T\mathcal{Q} = \{(q, v); q \in \mathcal{Q}, v \in T_q\mathcal{Q}\} = \{(q, v); g(q) = 0, G(q)v = 0\}, \quad (1.8)$$

the *tangent bundle* of \mathcal{Q} . Indeed, we have just shown that for initial values $(q_0, v_0) \in T\mathcal{Q}$ (i.e., consistent initial values) the problems (1.6) and (1.2) are equivalent, so that the solutions of (1.6) stay on $T\mathcal{Q}$.

Reversibility. The system (1.2) and the corresponding differential equation (1.6) are reversible with respect to the involution $\rho(q, v) = (q, -v)$, if $f(q, -v) = f(q, v)$. This follows at once from Example V.1.3, because the solution $\dot{v}(q, v)$ of (1.5) satisfies $\dot{v}(q, -v) = \dot{v}(q, v)$.

For the numerical solution of differential-algebraic equations “index reduction” is a very popular technique. This means that instead of directly treating the problem (1.2) one numerically solves the differential equation (1.6) on the manifold \mathcal{M} . Projection methods (Sect. IV.4) as well as methods based on local coordinates (Sect. IV.5) are much in use. If one is interested in a correct simulation of the reversible structure of the problem, the symmetric methods of Sect. V.4 can be applied. Here we do not repeat these approaches for this particular situation, instead we concentrate on the symplectic integration of constrained systems.

VII.1.2 Hamiltonian Formulation

In Sect. VI.1 we have seen that, for unconstrained mechanical systems, the equations of motion become more structured if we use the momentum coordinates $p = \frac{\partial L}{\partial \dot{q}} = M(q)\dot{q}$ in place of the velocity coordinates $v = \dot{q}$. Let us do the same for the constrained system (1.2). As in the proof of Theorem VI.1.3 we obtain the equivalent system

$$\begin{aligned} \dot{q} &= H_p(p, q) \\ \dot{p} &= -H_q(p, q) - G(q)^T \lambda \\ 0 &= g(q), \end{aligned} \quad (1.9)$$

where

$$H(p, q) = \frac{1}{2} p^T M(q)^{-1} p + U(q) \quad (1.10)$$

is the total energy of the system; H_p and H_q denote the column vectors of partial derivatives. Differentiating the constraint in (1.9) twice with respect to time, we get

$$0 = G(q)H_p(p, q), \quad (1.11)$$

$$0 = \frac{\partial}{\partial q} \left(G(q)H_p(p, q) \right) H_p(p, q) - G(q)H_{pp}(p, q) \left(H_q(p, q) + G(q)^T \lambda \right), \quad (1.12)$$

and assuming the matrix

$$G(q)H_{pp}(p, q)G(q)^T \quad \text{is invertible,} \quad (1.13)$$

equation (1.12) permits us to express λ in terms of (p, q) .

Formulation as a Differential Equation on a Manifold. Inserting the so-obtained function $\lambda(p, q)$ into (1.9) gives a differential equation for (p, q) on the manifold

$$\mathcal{M} = \{(p, q) ; g(q) = 0, G(q)H_p(p, q) = 0\}. \quad (1.14)$$

As we will now see, this manifold has a differential-geometric interpretation as the cotangent bundle of the configuration manifold $\mathcal{Q} = \{q ; g(q) = 0\}$. The Lagrangian for a fixed $q \in \mathcal{Q}$ is a function on the tangent space $T_q\mathcal{Q}$, i.e., $L(q, \cdot) : T_q\mathcal{Q} \rightarrow \mathbb{R}$. Its (Fréchet) derivative evaluated at $\dot{q} \in T_q\mathcal{Q}$ is therefore a linear mapping $d_{\dot{q}}L(q, \dot{q}) : T_q\mathcal{Q} \rightarrow \mathbb{R}$, or in other terms, $d_{\dot{q}}L(q, \dot{q})$ is in the cotangent space $T_q^*\mathcal{Q}$. Since the duality is such that $\langle d_{\dot{q}}L(q, \dot{q}), v \rangle = \frac{\partial L}{\partial \dot{q}}(q, \dot{q})v$ for $v \in T_q\mathcal{Q}$, condition (1.13) ensures that the Legendre transform $\dot{q} \mapsto p = d_{\dot{q}}L(q, \dot{q})$ is an invertible transformation between $T_q\mathcal{Q}$ and $T_q^*\mathcal{Q}$. We can therefore consider $T_q^*\mathcal{Q}$ as a subspace of \mathbb{R}^d if every $p \in T_q^*\mathcal{Q}$ is identified with $\frac{\partial L}{\partial \dot{q}}(q, \dot{q})^T = M(q)\dot{q} \in \mathbb{R}^d$ for the unique $\dot{q} \in T_q\mathcal{Q}$ for which $p = d_{\dot{q}}L(q, \dot{q})$ holds. With this identification,

$$T_q^*\mathcal{Q} = \{M(q)\dot{q} ; \dot{q} \in T_q\mathcal{Q}\},$$

and the duality is given by $\langle p, v \rangle = p^T v$ for $p \in T_q^*\mathcal{Q}$ and $v \in T_q\mathcal{Q}$. We thus have $p = M(q)\dot{q} \in T_q^*\mathcal{Q}$ if and only if $\dot{q} = M(q)^{-1}p = H_p(p, q) \in T_q\mathcal{Q}$. Since the tangent space at $q \in \mathcal{Q}$ is $T_q\mathcal{Q} = \{\dot{q} ; G(q)\dot{q} = 0\}$, we obtain that

$$p \in T_q^*\mathcal{Q} \quad \text{if and only if} \quad G(q)H_p(p, q) = 0.$$

Denoting by $T^*\mathcal{Q} = \{(p, q) ; q \in \mathcal{Q}, p \in T_q^*\mathcal{Q}\}$ the *cotangent bundle* of \mathcal{Q} , we thus see that the constraint manifold \mathcal{M} of (1.14) equals

$$\mathcal{M} = T^*\mathcal{Q}. \quad (1.15)$$

The constrained Hamiltonian system (1.9) with Hamiltonian (1.10) can thus be viewed as a differential equation on the cotangent bundle $T^*\mathcal{Q}$ of the configuration manifold \mathcal{Q} .

In the following we consider the system (1.9)–(1.12) with (1.13) where $H(p, q)$ is an arbitrary smooth function. The constraint manifold is then still given by (1.14). The existence and uniqueness of the solution of (1.9) can be discussed as before.

Reversibility. It is readily checked that the system (1.9) is reversible if $H(-p, q) = H(p, q)$. This is always satisfied for a Hamiltonian (1.10).

Preservation of the Hamiltonian. Differentiation of $H(p(t), q(t))$ with respect to time yields

$$-H_p^T H_q - H_p^T G^T \lambda + H_q^T H_p$$

with all expressions evaluated at $(p(t), q(t))$. The first and the last terms cancel, and the central term vanishes because $GH_p = 0$ on the solution manifold. Consequently, the Hamiltonian $H(p, q)$ is constant along solutions of (1.9).

Symplecticity of the Flow. Since the flow of the system (1.9) is a transformation on \mathcal{M} , its derivative is a mapping between the corresponding tangent spaces. In agreement with Definition VI.2.2 we call a map $\varphi : \mathcal{M} \rightarrow \mathcal{M}$ symplectic if, for every $x = (p, q) \in \mathcal{M}$,

$$\xi_1^T \varphi'(x)^T J \varphi'(x) \xi_2 = \xi_1^T J \xi_2 \quad \text{for all } \xi_1, \xi_2 \in T_x \mathcal{M}. \quad (1.16)$$

If φ is actually defined and continuously differentiable in an open subset of \mathbb{R}^{2d} that contains \mathcal{M} , then $\varphi'(x)$ in the above formula is just the usual Jacobian matrix. Otherwise, some care is necessary in the interpretation of (1.16): φ' is the tangent map given by the directional derivative $\varphi'(x)\xi := (d/d\tau)|_{\tau=0} \varphi(\gamma(\tau))$ for $\xi \in T_x \mathcal{M}$, where γ is a path on \mathcal{M} with $\gamma(0) = x$, $\dot{\gamma}(0) = \xi$. The expression $\xi_1^T \varphi'(x)^T$ in (1.16) should then be interpreted as $(\varphi'(x)\xi_1)^T$.

Theorem 1.2. *Let $H(p, q)$ and $g(q)$ be twice continuously differentiable. The flow $\varphi_t : \mathcal{M} \rightarrow \mathcal{M}$ of the system (1.9) is then a symplectic transformation on \mathcal{M} , i.e., it satisfies (1.16).*

Proof. We let $x = (p, q)$, so that the system (1.9) becomes $\dot{x} = J^{-1}(\nabla H(x) + \sum_i \lambda_i(x) \nabla g_i(x))$, where $\lambda_i(x)$ and $g_i(x)$ are the components of $\lambda(x)$ and $g(x)$, and $\lambda(x)$ is the function obtained from (1.12). The variational equation of this system, satisfied by the directional derivative $\dot{\Psi} = \varphi'_t(x_0)\xi$, with $x_0 = (p_0, q_0)$, reads

$$\dot{\Psi} = J^{-1} \left(\nabla^2 H(x) + \sum_{i=1}^m \lambda_i(x) \nabla^2 g_i(x) + \sum_{i=1}^m \nabla g_i(x) \nabla \lambda_i(x)^T \right) \Psi.$$

A direct computation, analogous to that in the proof of Theorem VI.2.4, yields for $\xi_1, \xi_2 \in T_{x_0} \mathcal{M}$

$$\begin{aligned} \frac{d}{dt} \left(\xi_1^T \varphi'_t(x_0)^T J \varphi'_t(x_0) \xi_2 \right) &= \dots = \sum_{i=1}^m \xi_1^T \varphi'_t(x_0)^T \nabla g_i(x) \nabla \lambda_i(x)^T \varphi'_t(x_0) \xi_2 \\ &\quad - \sum_{i=1}^m \xi_1^T \varphi'_t(x_0)^T \nabla \lambda_i(x) \nabla g_i(x)^T \varphi'_t(x_0) \xi_2. \end{aligned} \quad (1.17)$$

Since $g_i(\varphi_t(x_0)) = 0$ for $x_0 \in \mathcal{M}$, we have $\nabla g_i(x)^T \varphi'_t(x_0) \xi_2 = 0$ and the same for ξ_1 , so that the expression in (1.17) vanishes. This proves the symplecticity of the flow on \mathcal{M} . \square

Differentiating the constraint in (1.9) twice and solving for the Lagrange multiplier from (1.12) (this procedure is known as “index reduction” of the differential-algebraic system) yields the differential equation

$$\dot{q} = H_p(p, q), \quad \dot{p} = -H_q(p, q) - G(q)^T \lambda(p, q), \quad (1.18)$$

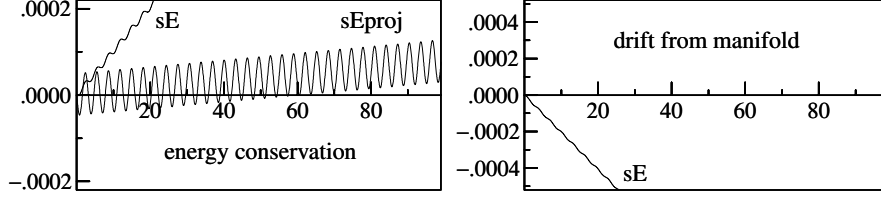


Fig. 1.1. Numerical solution of the symplectic Euler method applied to (1.18) with $H(p, q) = \frac{1}{2}(p_1^2 + p_2^2 + p_3^2) + q_3$, $g(q) = q_1^2 + q_2^2 + q_3^2 - 1$ (spherical pendulum); initial value $q_0 = (0, \sin(0.1), -\cos(0.1))$, $p_0 = (0.06, 0, 0)$, step size $h = 0.003$ for method “sE” (without projection) and $h = 0.03$ for method “sEproj” (with projection)

where $\lambda(p, q)$ is obtained from (1.12). If we solve this system with the symplectic Euler method (implicit in p , explicit in q), the qualitative behaviour of the numerical solution is not correct. As was observed by Leimkuhler & Reich (1994), there is a linear error growth in the Hamiltonian and also a drift from the manifold \mathcal{M} (method “sE” in Fig. 1.1). The explanation for this behaviour is the fact that (1.18) is no longer a Hamiltonian system. If we combine the symplectic Euler applied to (1.18) with an orthogonal projection onto \mathcal{M} (method “sEproj”), the result improves considerably but the linear error growth in the Hamiltonian is not eliminated. This numerical experiment illustrates that “index reduction” is not compatible with symplectic integration.

VII.1.3 A Symplectic First Order Method

We extend the symplectic Euler method to Hamiltonian systems with constraints. We integrate the p -variable by the implicit and the q -variable by the explicit Euler method. This gives

$$\begin{aligned}\hat{p}_{n+1} &= p_n - h (H_q(\hat{p}_{n+1}, q_n) + G(q_n)^T \lambda_{n+1}) \\ q_{n+1} &= q_n + h H_p(\hat{p}_{n+1}, q_n) \\ 0 &= g(q_{n+1}).\end{aligned}\tag{1.19}$$

The numerical approximation (\hat{p}_{n+1}, q_{n+1}) satisfies the constraint $g(q) = 0$, but not $G(q)H_p(p, q) = 0$. To get an approximation $(p_{n+1}, q_{n+1}) \in \mathcal{M}$, we append the projection

$$\begin{aligned}p_{n+1} &= \hat{p}_{n+1} - h G(q_{n+1})^T \mu_{n+1} \\ 0 &= G(q_{n+1})H_p(p_{n+1}, q_{n+1}).\end{aligned}\tag{1.20}$$

Let us discuss some basic properties of this method.

Existence and Uniqueness of the Numerical Solution. Inserting the definition of q_{n+1} from the second line of (1.19) into $0 = g(q_{n+1})$ gives a nonlinear system for \hat{p}_{n+1} and $h\lambda_{n+1}$. Due to the factor h in front of $H_p(\hat{p}_{n+1}, q_n)$, the implicit function theorem cannot be directly applied to prove existence and uniqueness of the numerical solution. We therefore write this equation as

$$0 = g(q_{n+1}) = g(q_n) + \int_0^1 G(q_n + \tau(q_{n+1} - q_n))(q_{n+1} - q_n) d\tau.$$

We now use $g(q_n) = 0$, insert the definition of q_{n+1} from the second line of (1.19) and divide by h . Together with the first line of (1.19) this yields the system $F(\hat{p}_{n+1}, h\lambda_{n+1}, h) = 0$ with

$$F(p, \nu, h) = \begin{pmatrix} p - p_n + hH_q(p, q_n) + G(q_n)^T \nu \\ \int_0^1 G(q_n + \tau h H_p(p, q_n)) H_p(p, q_n) d\tau \end{pmatrix}.$$

Since $(p_n, q_n) \in \mathcal{M}$ with \mathcal{M} from (1.14), we have $F(p_n, 0, 0) = 0$. Furthermore,

$$\frac{\partial F}{\partial(p, \nu)}(p_n, 0, 0) = \begin{pmatrix} I & G(q_n)^T \\ G(q_n)H_{pp}(p_n, q_n) & 0 \end{pmatrix},$$

and this matrix is invertible by (1.13). Consequently, an application of the implicit function theorem proves that the numerical solution $(\hat{p}_{n+1}, h\lambda_{n+1})$ (and hence also q_{n+1}) exists and is locally unique for sufficiently small h .

The projection step (1.20) constitutes a nonlinear system for p_{n+1} and $h\mu_{n+1}$, to which the implicit function theorem can be directly applied.

Convergence of Order 1. The above use of the implicit function theorem yields the rough estimates

$$\hat{p}_{n+1} = p_n + \mathcal{O}(h), \quad h\lambda_{n+1} = \mathcal{O}(h), \quad h\mu_{n+1} = \mathcal{O}(h),$$

which, together with the equations (1.19) and (1.20), give

$$q_{n+1} = q(t_{n+1}) + \mathcal{O}(h^2), \quad p_{n+1} = p(t_{n+1}) - G(q(t_{n+1}))^T \nu + \mathcal{O}(h^2),$$

where $(p(t), q(t))$ is the solution of (1.9) passing through $(p_n, q_n) \in \mathcal{M}$ at $t = t_n$. Inserting these relations into the second equation of (1.20) we get

$$0 = G(q(t))H_p(p(t), q(t)) + G(q(t))H_{pp}(p(t), q(t))G(q(t))^T \nu + \mathcal{O}(h^2)$$

at $t = t_{n+1}$. Since $G(q(t))H_p(p(t), q(t)) = 0$, it follows from (1.13) that $\nu = \mathcal{O}(h^2)$. The local error is therefore of size $\mathcal{O}(h^2)$.

The convergence proof now follows standard arguments, because the method is a mapping $\Phi_h : \mathcal{M} \rightarrow \mathcal{M}$ on the solution manifold. We consider the solutions $(p_n(t), q_n(t))$ of (1.9) passing through the numerical values $(p_n, q_n) \in \mathcal{M}$ at $t = t_n$, we estimate the difference of two successive solutions in terms of the local error at t_n , and we sum up the propagated errors (see Fig. 3.2 of Sect. II.3 in Hairer, Nørsett & Wanner (1993)). This proves that the global error satisfies $p_n - p(t_n) = \mathcal{O}(h)$ and $q_n - q(t_n) = \mathcal{O}(h)$ as long as $t_n = nh \leq \text{Const.}$

Symplecticity. We first study the mapping $(p_n, q_n) \mapsto (\hat{p}_{n+1}, q_{n+1})$ defined by (1.19), and we consider λ_{n+1} as a function $\lambda(p_n, q_n)$. Differentiation with respect to (p_n, q_n) yields

$$\begin{pmatrix} I + hH_{qp}^T & 0 \\ -hH_{pp} & I \end{pmatrix} \begin{pmatrix} \frac{\partial(\hat{p}_{n+1}, q_{n+1})}{\partial(p_n, q_n)} \end{pmatrix} = \begin{pmatrix} I - hG^T \lambda_p & S - hG^T \lambda_q \\ 0 & I + hH_{qp} \end{pmatrix}, \quad (1.21)$$

where $S = -hH_{qq} - h\lambda^T g_{qq}$ is a symmetric matrix, the expressions H_{qp} , H_{pp} , H_{qq} , G are evaluated at (\hat{p}_{n+1}, q_n) , and λ , λ_p , λ_q at (p_n, q_n) . A computation, identical to that of the proof of Theorem VI.3.3, yields

$$\left(\frac{\partial(\hat{p}_{n+1}, q_{n+1})}{\partial(p_n, q_n)} \right)^T J \left(\frac{\partial(\hat{p}_{n+1}, q_{n+1})}{\partial(p_n, q_n)} \right) = \begin{pmatrix} 0 & I - h\lambda_p^T G \\ -I + hG^T \lambda_p & h(G^T \lambda_q - \lambda_q^T G) \end{pmatrix}.$$

We multiply this relation from the left by $\xi_1 \in T_{(p_n, q_n)}\mathcal{M}$ and from the right by $\xi_2 \in T_{(p_n, q_n)}\mathcal{M}$. With the partitioning $\xi = (\xi_p, \xi_q)$ we have $G(q_n)\xi_{q,j} = 0$ for $j = 1, 2$ so that the expression reduces to $\xi_1^T J \xi_2$. This proves the symplecticity condition (1.16) for the mapping $(p_n, q_n) \mapsto (\hat{p}_{n+1}, q_{n+1})$.

Similarly, the projection step $(\hat{p}_{n+1}, q_{n+1}) \mapsto (p_{n+1}, q_{n+1})$ of (1.20) gives

$$\frac{\partial(p_{n+1}, q_{n+1})}{\partial(\hat{p}_{n+1}, q_{n+1})} = \begin{pmatrix} I - hG^T \mu_p & S - hG^T \mu_q \\ 0 & I \end{pmatrix},$$

where μ_{n+1} of (1.20) is considered as a function of (\hat{p}_{n+1}, q_{n+1}) , and $S = -h\mu^T g_{qq}$. This is formally the same as (1.21) with $H \equiv 0$. Consequently, the symplecticity condition is also satisfied for this mapping. As a composition of two symplectic transformations, the numerical flow of our first order method is therefore also symplectic.

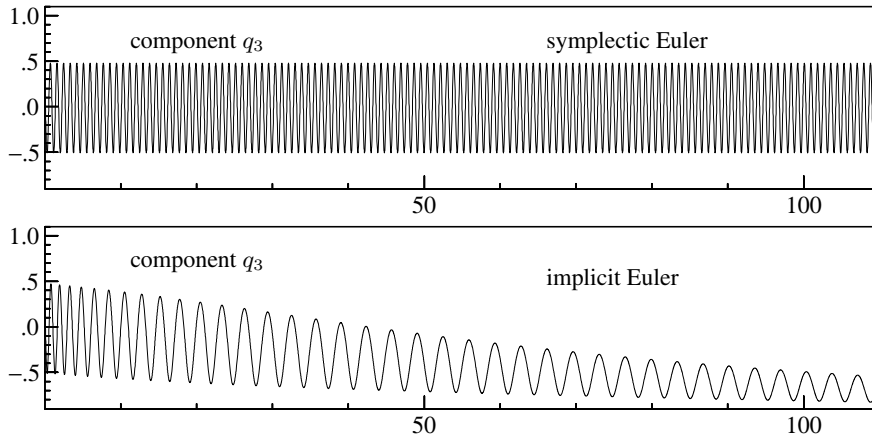


Fig. 1.2. Spherical pendulum problem solved with the symplectic Euler method (1.19)-(1.20) and with the implicit Euler method; initial value $q_0 = (\sin(1.3), 0, \cos(1.3))$, $p_0 = (3 \cos(1.3), 6.5, -3 \sin(1.3))$, step size $h = 0.01$

Numerical Experiment. Consider the equations (1.3) for the spherical pendulum. For a mass $m = 1$ they coincide with the Hamiltonian formulation. Figure 1.2 (upper picture) shows the numerical solution (vertical coordinate q_3) over many periods obtained by method (1.19)-(1.20). We observe a regular qualitatively correct behaviour. For the implicit Euler method (i.e., the argument q_n is replaced with q_{n+1} in (1.19)) the numerical solution, obtained with the same step size and the same initial values, is less satisfactory. Already after one period the solution deteriorates and the pendulum loses energy.

VII.1.4 SHAKE and RATTLE

The numerical method (1.19)-(1.20) is only of order 1 and it is not symmetric. An algorithm that is of order 2, symmetric and symplectic was originally considered for separable Hamiltonians

$$H(p, q) = \frac{1}{2} p^T M^{-1} p + U(q) \quad (1.22)$$

with constant mass matrix M . Notice that in this case we are concerned with a second order differential equation $M\ddot{q} = -U_q(q) - G(q)^T \lambda$ with $g(q) = 0$.

SHAKE. Ryckaert, Ciccotti & Berendsen (1977) propose the method

$$\begin{aligned} q_{n+1} - 2q_n + q_{n-1} &= -h^2 M^{-1} (U_q(q_n) + G(q_n)^T \lambda_n) \\ 0 &= g(q_{n+1}) \end{aligned} \quad (1.23)$$

for computations in molecular dynamics. It is a straightforward extension of the Störmer-Verlet scheme (I.1.15). The p -components, not used in the recursion, are approximated by $p_n = M(q_{n+1} - q_{n-1})/2h$.

RATTLE. The three-term recursion (1.23) may lead to an accumulation of round-off errors, and a reformulation as a one-step method is desirable. Using the same procedure as in (I.1.17) we formally get

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2} (U_q(q_n) + G(q_n)^T \lambda_n) \\ q_{n+1} &= q_n + h M^{-1} p_{n+1/2}, \quad 0 = g(q_{n+1}) \\ p_{n+1} &= p_{n+1/2} - \frac{h}{2} (U_q(q_{n+1}) + G(q_{n+1})^T \lambda_{n+1}). \end{aligned} \quad (1.24)$$

The difficulty with this formulation is that λ_{n+1} is not yet available at this step (it is computed together with q_{n+2}). As a remedy, Andersen (1983) suggests replacing the last line in (1.24) with a projection step similar to (1.20)

$$\begin{aligned} p_{n+1} &= p_{n+1/2} - \frac{h}{2} (U_q(q_{n+1}) + G(q_{n+1})^T \mu_n) \\ 0 &= G(q_{n+1}) M^{-1} p_{n+1}. \end{aligned} \quad (1.25)$$

This modification, called RATTLE, has the further advantage that the numerical approximation (p_{n+1}, q_{n+1}) lies on the solution manifold \mathcal{M} . The symplecticity of this algorithm has been established by Leimkuhler & Skeel (1994).

Extension to General Hamiltonians. As observed independently by Jay (1994) and Reich (1993), the RATTLE algorithm can be extended to general Hamiltonians as follows: for consistent values $(p_n, q_n) \in \mathcal{M}$ define

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2} (H_q(p_{n+1/2}, q_n) + G(q_n)^T \lambda_n) \\ q_{n+1} &= q_n + \frac{h}{2} (H_p(p_{n+1/2}, q_n) + H_p(p_{n+1/2}, q_{n+1})) \\ 0 &= g(q_{n+1}) \\ p_{n+1} &= p_{n+1/2} - \frac{h}{2} (H_q(p_{n+1/2}, q_{n+1}) + G(q_{n+1})^T \mu_n) \\ 0 &= G(q_{n+1}) H_p(p_{n+1}, q_{n+1}). \end{aligned} \tag{1.26}$$

The first three equations of (1.26) are very similar to (1.19) and the last two equations to (1.20). The existence of (locally) unique solutions $(p_{n+1/2}, q_{n+1}, \lambda_n)$ and (p_{n+1}, μ_n) can therefore be proved in the same way. Notice also that this method gives a numerical solution that stays exactly on the solution manifold \mathcal{M} .

Theorem 1.3. *The numerical method (1.26) is symmetric, symplectic, and convergent of order two.*

Proof. Although this theorem is the special case $s = 2$ of Theorem 1.4, we outline its proof. We will see that the convergence result is easier to obtain for $s = 2$ than for the general case.

If we add to (1.26) the consistency conditions $g(q_n) = 0$, $G(q_n) H_p(p_n, q_n) = 0$ of the initial values, the symmetry of the method follows at once by exchanging $h \leftrightarrow -h$, $p_{n+1} \leftrightarrow p_n$, $q_{n+1} \leftrightarrow q_n$, and $\lambda_n \leftrightarrow \mu_n$. The symplecticity can be proved as for (1.19)-(1.20) by computing the derivative of (p_{n+1}, q_{n+1}) with respect to (p_n, q_n) , and by verifying the condition (1.16). This does not seem to be simpler than the symplecticity proof of Theorem 1.4.

The implicit function theorem applied to the two subsystems of (1.26) shows

$$p_{n+1/2} = p_n + \mathcal{O}(h), \quad h\lambda = \mathcal{O}(h), \quad p_{n+1} = p_{n+1/2} + \mathcal{O}(h), \quad h\mu = \mathcal{O}(h),$$

and, inserted into (1.26), yields

$$q_{n+1} = q(t_{n+1}) + \mathcal{O}(h^2), \quad p_{n+1} = p(t_{n+1}) - G(q(t_{n+1}))^T \nu + \mathcal{O}(h^2).$$

Convergence of order one follows therefore in the same way as for method (1.19)-(1.20). Since the order of a symmetric method is always even, this implies convergence of order two. \square

An easy way of obtaining high order methods for constrained Hamiltonian systems is by composition (Reich 1996a). Method (1.26) is an ideal candidate as basic integrator for compositions of the form (V.3.2). The resulting integrators are symmetric, symplectic, of high order, and yield a numerical solution that stays on the manifold \mathcal{M} .

VII.1.5 The Lobatto IIIA - IIIB Pair

Another possibility for obtaining high order symplectic integrators for constrained Hamiltonian systems is by the use of partitioned Runge–Kutta or discontinuous collocation methods. We consider the system (1.9) and we search for polynomials $u(t)$ of degree s , $w(t)$ of degree $s - 1$, and $v(t)$ of degree $s - 2$ such that

$$u(t_n) = q_n, \quad v(t_n) = p_n - hb_1\delta(t_n) \quad (1.27)$$

with the defect

$$\delta(t) = \dot{v}(t) + H_q(v(t), u(t)) + G(u(t))^T w(t) \quad (1.28)$$

and, using the abbreviation $t_{n,i} = t_n + c_i h$,

$$\dot{u}(t_{n,i}) = H_p(v(t_{n,i}), u(t_{n,i})), \quad i = 1, \dots, s \quad (1.29)$$

$$\dot{v}(t_{n,i}) = -H_q(v(t_{n,i}), u(t_{n,i})) - G(u(t_{n,i}))^T w(t_{n,i}), \quad i = 2, \dots, s - 1$$

$$0 = g(u(t_{n,i})), \quad i = 1, \dots, s.$$

If these polynomials exist, the numerical solution is defined by

$$\begin{aligned} q_{n+1} &= u(t_n + h), & p_{n+1} &= v(t_n + h) - hb_s\delta(t_n + h) \\ 0 &= G(q_{n+1})H_p(p_{n+1}, q_{n+1}). \end{aligned} \quad (1.30)$$

Why Discontinuous Collocation Based on Lobatto Quadrature? At a first glance (Theorem VI.4.2) it seems natural to consider collocation methods based on Gaussian quadrature for the entire system. This, however, has the disadvantage that the numerical solution does not satisfy $g(q_{n+1}) = 0$. To achieve this requirement, $t_n + h$ has to be one of the collocation points, i.e., we must have $c_s = 1$. Unfortunately, none of the collocation or discontinuous collocation methods with $c_s = 1$ is symplectic (see Exercise IV.6). We therefore turn our attention to partitioned methods, and we treat only the q -component by a collocation method satisfying $c_s = 1$. To satisfy the s conditions $g(u(t_{n,i})) = 0$ of (1.29) there are only $s - 1$ free parameters $w(t_n), w(t_n + c_2 h), \dots, w(t_n + c_{s-1} h)$ available. A remedy is to choose $c_1 = 0$ so that the first condition $g(u(t_n)) = 0$ is automatically verified. Encouraged by Theorem VI.4.5 we are thus led to consider the Lobatto nodes in the role of the c_i . The use of the partitioned Lobatto IIIA - IIIB pair for the treatment of constrained Hamiltonian systems has been suggested by Jay (1994, 1996).

Existence and Uniqueness of the Numerical Solution. The polynomial $u(t)$ of degree s is uniquely determined by $u(t_n) = q_n$ and $\dot{u}(t_{n,i}) =: \dot{Q}_i$ ($i = 1, \dots, s$), the polynomial $v(t)$ of degree $s - 2$ is uniquely determined by $v(t_{n,i}) =: P_i$ ($i = 1, \dots, s - 1$), and the polynomial $w(t)$ of degree $s - 1$ is uniquely determined by $hw(t_{n,i}) =: A_i$ ($i = 1, \dots, s$). Notice that the value A_s is only involved in (1.30) and not in (1.27)–(1.29). For the nonlinear system (1.27)–(1.29) we therefore consider

$$X = (\dot{Q}_1, \dots, \dot{Q}_s, P_1, \dots, P_{s-1}, A_1, \dots, A_{s-1})$$

as independent variables, and we write the system as $F(X, h) = 0$. The function F is composed of the s conditions for $\dot{u}(t_{n,i})$, of the definition of $v(t_n)$ (divided by b_1) and the $s - 2$ conditions for $\dot{v}(t_{n,i})$ (multiplied by h), and finally of the $s - 1$ equations $0 = g(u(t_{n,i}))$ for $i = 2, \dots, s$ (divided by h). Observe that $0 = g(u(t_n))$ is automatically satisfied by the consistency of (p_n, q_n) . We note that $P_s = v(t_n + h)$ and $\dot{P}_i = h\dot{v}(t_{n,i})$ are linear combinations of P_1, \dots, P_{s-1} with coefficients independent of the step size h .

The function $F(X, h)$ is well-defined for h in a neighbourhood of 0. For the first two blocks this is evident, for the last one it follows from the identity

$$\frac{1}{h} g(u(t_{n,i})) = \int_0^{c_i} G(u(t_n + \theta h)) \dot{u}(t_n + \theta h) d\theta$$

using the fact that $\dot{u}(t_n + \theta h)$ is a linear combination of \dot{Q}_i for $i = 1, \dots, s$. With the values

$$X_0 = (H_p(p_n, q_n), \dots, H_p(p_n, q_n), p_n, \dots, p_n, 0, \dots, 0)$$

we have that $F(X_0, 0) = 0$, because the values (p_n, q_n) are assumed to be consistent. In view of an application of the implicit function theorem we compute

$$\frac{\partial F}{\partial X}(X_0, 0) = \begin{pmatrix} I \otimes I & -D \otimes H_{pp} & 0 \\ 0 & B \otimes I & I \otimes G^T \\ A \otimes G & 0 & 0 \end{pmatrix}, \quad (1.31)$$

where H_{pp} , G are evaluated at (p_n, q_n) , and A, B, D are matrices of dimension $(s - 1) \times s$, $(s - 1) \times (s - 1)$ and $s \times (s - 1)$ respectively that depend only on the Lobatto quadrature and not on the differential equation. For example, the matrix B represents the linear mapping

$$(P_1, \dots, P_{s-1}) \mapsto (\dot{P}_1 + b_1^{-1} P_1, \dot{P}_2, \dots, \dot{P}_{s-1}).$$

This mapping is invertible, because the values on the right-hand side uniquely determine the polynomial $v(t)$ of degree $s - 2$.

Block Gaussian elimination then shows that (1.31) is invertible if and only if the matrix

$$ADB^{-1} \otimes GH_{pp}G^T \quad \text{is invertible.}$$

Because of (1.13) it remains to show that ADB^{-1} is invertible.

To achieve this without explicitly computing the matrices A, B, D , we apply the method to the problem where p and q are of dimension one, $H(p, q) = p^2/2$, and $g(q) = q$. Assuming $h = 1$ we get

$$\begin{aligned} u(0) &= 0, & v(0) &= -b_1(\dot{v}(0) + w(0)) \\ \dot{u}(c_i) &= v(c_i) & \text{for } i &= 1, \dots, s \\ \dot{v}(c_i) &= -w(c_i) & \text{for } i &= 2, \dots, s - 1 \\ 0 &= u(c_i) & \text{for } i &= 1, \dots, s, \end{aligned} \quad (1.32)$$

which is equivalent to

$$\begin{pmatrix} I & -D & 0 \\ 0 & B & I \\ A & 0 & 0 \end{pmatrix} \begin{pmatrix} (\dot{u}(c_i))_{i=1}^s \\ (v(c_i))_{i=1}^{s-1} \\ (w(c_i))_{i=1}^{s-1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad (1.33)$$

because $H_{pp}(p, q) = 1$ and $G(q) = 1$. Since $u(t)$ is a polynomial of degree s , the last equation of (1.32) implies that $u(t) = C \prod_{j=1}^s (t - c_j)$. By the second relation the polynomial $\dot{u}(t) - v(t)$, which is of degree $s - 1$, vanishes at s points. Hence, $v(t) \equiv \dot{u}(t)$, which is possible only if $C = 0$, because the degree of $v(t)$ is $s - 2$. Consequently, the linear system (1.33) has only the trivial solution, so that the matrix in (1.33) and hence also ADB^{-1} is invertible.

The implicit function theorem applied to $F(X, h) = 0$ shows that the nonlinear system (1.27)-(1.30) possesses a locally unique solution for sufficiently small step sizes h . Using the free parameter $\Lambda_s = hw(t_n + h)$, a further application of the implicit function theorem, this time to the small system (1.30), proves the existence and local uniqueness of p_{n+1} .

Theorem 1.4. *Let $(b_i, c_i)_{i=1}^s$ be the weights and nodes of the Lobatto quadrature (c.f. (II.1.17)). The method (1.27)-(1.29)-(1.30) is symmetric, symplectic, and superconvergent of order $2s - 2$.*

Proof. Symmetry. To the formulas (1.27)-(1.29)-(1.30) we add the consistency relations $g(q_n) = 0$, $G(q_n)H_p(p_n, q_n) = 0$. Then we exchange $(t_n, p_n, q_n) \leftrightarrow (t_{n+1}, p_{n+1}, q_{n+1})$ and $h \leftrightarrow -h$. Since $b_1 = b_s$ and $c_{s+1-i} = 1 - c_i$ for the Lobatto quadrature, the resulting formulas are equivalent to the original method (see also the proof of Theorem V.2.1).

Symplecticity. We fix $\xi_1, \xi_2 \in T_{(p_n, q_n)}\mathcal{M}$, we put $x_n = (p_n, q_n)^T$, and we consider the bilinear mapping

$$Q\left(\frac{\partial p_{n+1}}{\partial x_n}, \frac{\partial q_{n+1}}{\partial x_n}\right) = \xi_1^T \left(\left(\frac{\partial q_{n+1}}{\partial x_n} \right)^T \left(\frac{\partial p_{n+1}}{\partial x_n} \right) - \left(\frac{\partial p_{n+1}}{\partial x_n} \right)^T \left(\frac{\partial q_{n+1}}{\partial x_n} \right) \right) \xi_2.$$

The symplecticity of the transformation $(p_n, q_n) \mapsto (p_{n+1}, q_{n+1})$ on the manifold \mathcal{M} is then expressed by the relation

$$Q\left(\frac{\partial p_{n+1}}{\partial x_n}, \frac{\partial q_{n+1}}{\partial x_n}\right) = Q\left(\frac{\partial p_n}{\partial x_n}, \frac{\partial q_n}{\partial x_n}\right). \quad (1.34)$$

We now follow closely the proof of Theorem IV.2.3. We consider the polynomials $u(t), v(t), w(t)$ of the method (1.27)-(1.29)-(1.30) as functions of t and $x_n = (p_n, q_n)$, and we compute

$$\begin{aligned} Q\left(\frac{\partial v(t_{n+1})}{\partial x_n}, \frac{\partial u(t_{n+1})}{\partial x_n}\right) &= Q\left(\frac{\partial v(t_n)}{\partial x_n}, \frac{\partial u(t_n)}{\partial x_n}\right) \\ &= \int_{t_n}^{t_{n+1}} \frac{dQ}{dt} \left(\frac{\partial v(t)}{\partial x_n}, \frac{\partial u(t)}{\partial x_n} \right) dt. \end{aligned} \quad (1.35)$$

Since $u(t)$ is a polynomial of degree s and $v(t)$ of degree $s - 2$, the integrand in (1.35) is a polynomial in t of degree $2s - 3$. It is thus integrated without error by the Lobatto quadrature. By definition these polynomials satisfy the differential equation at the interior collocation points. Therefore, it follows from (1.17) that

$$\frac{dQ}{dt} \left(\frac{\partial v(t_{n,i})}{\partial x_n}, \frac{\partial u(t_{n,i})}{\partial x_n} \right) = 0 \quad \text{for } i = 2, \dots, s-1,$$

and that

$$\frac{dQ}{dt} \left(\frac{\partial v(t_{n,i})}{\partial x_n}, \frac{\partial u(t_{n,i})}{\partial x_n} \right) = Q \left(\frac{\partial \delta(t_{n,i})}{\partial x_n}, \frac{\partial u(t_{n,i})}{\partial x_n} \right) \quad \text{for } i = 1 \text{ and } i = s.$$

Applying the Lobatto quadrature to the integral in (1.35) thus yields

$$hb_1 Q \left(\frac{\partial \delta(t_n)}{\partial x_n}, \frac{\partial u(t_n)}{\partial x_n} \right) + hb_s Q \left(\frac{\partial \delta(t_{n+1})}{\partial x_n}, \frac{\partial u(t_{n+1})}{\partial x_n} \right),$$

and the symplecticity relation (1.34) follows in the same way as in the proof of Theorem IV.2.3.

Superconvergence. This is the most difficult part of the proof. We remark that superconvergence of Runge–Kutta methods for differential-algebraic systems of index 3 has been conjectured by Hairer, Lubich & Roche (1989), and a first proof has been obtained by Jay (1993) for collocation methods. In his thesis Jay (1994) proves superconvergence for a more general class of methods, including the Lobatto IIIA - IIIB pair, using a “rooted-tree-type” theory. A sketch of that very elaborate proof is published in Jay (1996). Using the idea of discontinuous collocation, the elegant proof for collocation methods can now be extended to cover the Lobatto IIIA - IIIB pair. In the following we explain how the local error can be estimated.

We consider the polynomials $u(t), v(t), w(t)$ defined in (1.27)-(1.29)-(1.30), and we define defects $\mu(t), \delta(t), \theta(t)$ as follows:

$$\begin{aligned} \dot{u}(t) &= H_p(v(t), u(t)) + \mu(t) \\ \dot{v}(t) &= -H_q(v(t), u(t)) - G(u(t))^T w(t) + \delta(t) \\ 0 &= g(u(t)) + \theta(t). \end{aligned} \tag{1.36}$$

By definition of the method we have

$$\begin{aligned} \mu(t_n + c_i h) &= 0, & i = 1, \dots, s \\ \delta(t_n + c_i h) &= 0, & i = 2, \dots, s-1 \\ \theta(t_n + c_i h) &= 0, & i = 1, \dots, s. \end{aligned} \tag{1.37}$$

We let $q(t), p(t), \lambda(t)$ be the exact solution of (1.9) satisfying $q(t_n) = q_n, p(t_n) = p_n$, and we consider the differences

$$\Delta u(t) = u(t) - q(t), \quad \Delta v(t) = v(t) - p(t), \quad \Delta w(t) = w(t) - \lambda(t).$$

Subtracting (1.9) from (1.36) we get by linearization that

$$\begin{aligned}\dot{\Delta u} &= a_{11}(t)\Delta u + a_{12}(t)\Delta v + \mu(t) \\ \dot{\Delta v} &= a_{21}(t)\Delta u + a_{22}(t)\Delta v + a_{23}(t)\Delta w + \delta(t),\end{aligned}\tag{1.38}$$

where $a_{12}(t) = H_{pp}(p(t), q(t))$, and where the other $a_{ij}(t)$ are given by similar expressions. We have suppressed quadratic and higher order terms to keep the presentation as simple as possible. They do not influence the convergence result. To eliminate Δw in (1.38), we differentiate the algebraic relations in (1.9) and (1.36) twice, and we subtract them. This yields

$$\begin{aligned}0 &= F(t, \mu(t)) + b_1(t)\Delta u + b_2(t)\Delta v + B(t)\Delta w \\ &+ G(u(t))H_{pp}(v(t), u(t))\delta(t) + G(u(t))\dot{\mu}(t) + \ddot{\theta}(t),\end{aligned}$$

where $F(t, \mu)$, $B(t)$, $b_1(t)$, $b_2(t)$ are functions depending on $p(t)$, $q(t)$, $\lambda(t)$, $u(t)$, $v(t)$, $w(t)$, and where $F(t, 0) = 0$ and $B(t) \approx G(q_n)H_{pp}(p_n, q_n)G(q_n)^T$. Because of our assumption (1.13) we can extract Δw from this relation, and we insert it into (1.38). In this way we get a linear differential equation for Δu , Δv , which can be solved by the “variation of constants” formula. Using $\Delta u(t_n) = 0$ (by (1.27)), the solution $\Delta v(t_n + h)$ is seen to be of the form

$$\begin{aligned}\Delta v(t_n + h) &= R_{22}(t_n + h, t_n)\Delta v(t_n) + \int_{t_n}^{t_n+h} \left(R_{21}(t_n + h, t)\mu(t) \right. \\ &+ R_{22}(t_n + h, t) \left(\delta(t) + \tilde{F}(t, \mu(t)) + c_1(t)\dot{\mu}(t) \right. \\ &\left. \left. + C(t) \left(G(u(t))H_{pp}(v(t), u(t))\delta(t) + \ddot{\theta}(t) \right) \right) \right) dt,\end{aligned}\tag{1.39}$$

where R_{21} and R_{22} are the lower blocks of the resolvent, and \tilde{F} , c_1 , C are functions as before. To prove that the local error of the p -component

$$p_{n+1} - p(t_n + h) = \Delta v(t_n + h) - hb_s\delta(t_n + h)\tag{1.40}$$

is of size $\mathcal{O}(h^{2s-1})$, we first integrate by parts those expressions in (1.39) which contain a derivative. For example,

$$\int_{t_n}^{t_{n+1}} a(t)\dot{\mu}(t) dt = a(t)\mu(t) \Big|_{t_n}^{t_{n+1}} - \int_{t_n}^{t_{n+1}} \dot{a}(t)\mu(t) dt = \mathcal{O}(h^{2s-1}),$$

because $\mu(t_n) = \mu(t_n + h) = 0$ by (1.37) and an application of the Lobatto quadrature to the integral at the right-hand side gives zero as result with a quadrature error of size $\mathcal{O}(h^{2s-1})$. Similarly, integrating by parts twice yields

$$\begin{aligned}\int_{t_n}^{t_{n+1}} a(t)\ddot{\theta}(t) dt &= a(t)\dot{\theta}(t) \Big|_{t_n}^{t_{n+1}} - \dot{a}(t)\theta(t) \Big|_{t_n}^{t_{n+1}} + \int_{t_n}^{t_{n+1}} \ddot{a}(t)\theta(t) dt \\ &= a(t_{n+1})\dot{\theta}(t_{n+1}) - a(t_n)\dot{\theta}(t_n) + \mathcal{O}(h^{2s-1}).\end{aligned}$$

To the other integrals in (1.39) we apply the Lobatto quadrature directly. Since $R_{22}(t_{n+1}, t_{n+1})$ is the identity, this gives

$$\begin{aligned} p_{n+1} - p(t_{n+1}) &= R_{22}(t_{n+1}, t_n) \left(\Delta v(t_n) + hb_1 \delta(t_n) \right) \\ &+ \tilde{C}(t_{n+1}) \left(hb_s G(u(t_{n+1})) H_{pp}(v(t_{n+1}), u(t_{n+1})) \delta(t_{n+1}) + \dot{\theta}(t_{n+1}) \right) \\ &+ \tilde{C}(t_n) \left(hb_1 G(u(t_n)) H_{pp}(v(t_n), u(t_n)) \delta(t_n) - \dot{\theta}(t_n) \right) + \mathcal{O}(h^{2s-1}), \end{aligned} \quad (1.41)$$

where $\tilde{C}(t) = R(t_{n+1}, t)C(t)$. The term $\Delta v(t_n) + hb_1 \delta(t_n)$ vanishes by (1.27), and differentiation of the algebraic relation in (1.36) yields

$$0 = G(u(t)) \left(H_p(v(t), u(t)) + \mu(t) \right) + \dot{\theta}(t).$$

As a consequence of (1.27), (1.37) and the consistency of the initial values (p_n, q_n) , this gives

$$\begin{aligned} \dot{\theta}(t_n) &= -G(q_n) H_p(p_n - hb_1 \delta(t_n), q_n) \\ &= hb_1 G(q_n) H_{pp}(p_n, q_n) \delta(t_n) + \mathcal{O}(h^2 \delta(t_n)^2) \\ &= hb_1 G(u(t_n)) H_{pp}(v(t_n), u(t_n)) \delta(t_n) + \mathcal{O}(h^2 \delta(t_n)^2). \end{aligned}$$

Using (1.30) we get in the same way

$$\dot{\theta}(t_{n+1}) = -hb_s G(u(t_{n+1})) H_{pp}(v(t_{n+1}), u(t_{n+1})) \delta(t_{n+1}) + \mathcal{O}(h^2 \delta(t_{n+1})^2).$$

These estimates together show that the local error (1.41) is of size $\mathcal{O}(h^{2s-1}) + \mathcal{O}(h^2 \delta(t)^2)$. The defect $\delta(t)$ vanishes at $s - 2$ points in the interval $[t_n, t_{n+1}]$, so that $\delta(t) = \mathcal{O}(h^{s-2})$ for $t \in [t_n, t_{n+1}]$ (for a rigorous proof of this statement one has to apply the techniques of the proof of Theorem II.1.5). Therefore we obtain $p_{n+1} - p(t_{n+1}) = \mathcal{O}(h^{2s-2})$, and by the symmetry of the method also $\mathcal{O}(h^{2s-1})$.

In analogy to (1.39), the variation of constants formula yields also an expression for the local error $q_{n+1} - q(t_{n+1}) = \Delta u(t_{n+1})$. One only has to replace R_{21} and R_{22} with the upper blocks R_{11} and R_{12} of the resolvent. Using $R_{12}(t_{n+1}, t_{n+1}) = 0$, we prove in the same way that the local error of the q -component is of size $\mathcal{O}(h^{2s-1})$.

The estimation of the global error is obtained in the same way as for the first order method (1.19)-(1.20). Since the algorithm is a mapping $\Phi_h : \mathcal{M} \rightarrow \mathcal{M}$ on the solution manifold, it is not necessary to follow the technically difficult proofs in the context of differential-algebraic equations. Summing up the propagated local errors proves that the global error satisfies $p_n - p(t_n) = \mathcal{O}(h^{2s-2})$ and $q_n - q(t_n) = \mathcal{O}(h^{2s-2})$ as long as $t_n = nh \leq \text{Const.}$ \square

VII.1.6 Splitting Methods

When considering splitting methods for constrained mechanical systems, it should be borne in mind that such systems are differential equations on manifolds (see

Sect. VII.1.2). Splitting methods should therefore be based on a decomposition $f(y) = f^{[1]}(y) + f^{[2]}(y)$, where both $f^{[i]}(y)$ are vector fields on the same manifold as $f(y)$. Let us consider here the Hamiltonian system (1.9) with Hamiltonian

$$H(p, q) = H^{[1]}(p, q) + H^{[2]}(p, q). \quad (1.42)$$

The manifold for this differential equation is

$$\mathcal{M} = \{(p, q) \mid g(q) = 0, G(q)H_p(p, q) = 0\}. \quad (1.43)$$

Notice that (1.9), when H is simply replaced with $H^{[i]}$, is not a good candidate for splitting methods: the existence of a solution is not guaranteed, and if the solution exists it need not stay on the manifold \mathcal{M} . The following lemma indicates how splitting methods should be applied.

Lemma 1.5. *Consider a Hamiltonian (1.42), a function $g(q)$ with $G(q) = g'(q)$, and let the manifold \mathcal{M} be given by (1.43). If (1.13) holds and if*

$$G(q)H_p^{[i]}(p, q) = 0 \quad \text{for all } (p, q) \in \mathcal{M}, \quad (1.44)$$

then the system

$$\begin{aligned} \dot{q} &= H_p^{[i]}(p, q) \\ \dot{p} &= -H_q^{[i]}(p, q) - G(q)^T \lambda \\ 0 &= G(q)H_p(p, q) \end{aligned} \quad (1.45)$$

defines a differential equation on the manifold \mathcal{M} , and its flow is a symplectic transformation on \mathcal{M} .

Proof. Differentiation of the algebraic relation in (1.45) with respect to time, and replacing \dot{q} and \dot{p} with their differential equations, yields an explicit relation for $\lambda = \lambda(p, q)$ (as a consequence of (1.13)). Hence, a unique solution of (1.45) exists locally if $G(q_0)H_p(p_0, q_0) = 0$. The assumption (1.44) implies $\frac{d}{dt}g(q(t)) = 0$. This together with the algebraic relation of (1.45) guarantees that for $(p_0, q_0) \in \mathcal{M}$ the solution stays on the manifold \mathcal{M} . The symplecticity of the flow is proved as for Theorem 1.2. \square

Suppose now that the Hamiltonian $H(p, q)$ of (1.9) can be split as in (1.42), where both $H^{[i]}(p, q)$ satisfy (1.44). We denote by $\varphi_t^{[i]}$ the flow of the system (1.45). If these flows can be computed analytically, the Lie-Trotter splitting $\varphi_h^{[2]} \circ \varphi_h^{[1]}$ and the Strang splitting $\varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}$ yield first and second order numerical integrators, respectively. Considering more general compositions as in (II.5.6) and using the coefficients proposed in Sect. V.3, methods of high order are obtained. They give numerical approximations lying on the manifold \mathcal{M} , and they are symplectic (also symmetric if the splitting is well chosen).

For the important special case where

$$H(p, q) = T(p, q) + U(q)$$

is the sum of the kinetic and potential energies, both summands satisfy assumption (1.44). This gives a natural splitting that is often used in practice.

Example 1.6 (Spherical Pendulum). We normalize all constants to 1 (cf. Example 1.1) and we consider the problem (1.9) with

$$H(p, q) = \frac{1}{2}(p_1^2 + p_2^2 + p_3^2) + q_3, \quad g(q) = \frac{1}{2}(q_1^2 + q_2^2 + q_3^2 - 1).$$

We split the Hamiltonian as $H^{[1]}(p, q) = \frac{1}{2}(p_1^2 + p_2^2 + p_3^2)$ and $H^{[2]}(p, q) = q_3$, and we solve (1.45) with initial values on the manifold

$$\mathcal{M} = \{(p, q) \mid q_1^2 + q_2^2 + q_3^2 - 1 = 0, p_1 q_1 + p_2 q_2 + p_3 q_3 = 0\}.$$

The kinetic energy $H^{[1]}(p, q)$ leads to the system

$$\dot{q} = p, \quad \dot{p} = -q\lambda, \quad q^T p = 0,$$

which gives $\lambda = p_0^T p_0$, so that the flow $\varphi_t^{[1]}$ is just a planar rotation around the origin. The potential energy $H^{[2]}(p, q)$ leads to

$$\dot{q} = 0, \quad \dot{p} = -(0, 0, 1)^T - q\lambda, \quad q^T p = 0.$$

The flow $\varphi_t^{[2]}$ keeps $q(t)$ constant and changes $p(t)$ linearly with time. Splitting methods give simple, explicit and symplectic time integrators for this problem.

VII.2 Poisson Systems

This section is devoted to an interesting generalization of Hamiltonian systems, where J^{-1} in (VI.2.5) is replaced with a nonconstant matrix $B(y)$. Such structures were introduced by Sophus Lie (1888) and are today called *Poisson systems*. They result, in particular, from Hamiltonian systems on manifolds written in non-canonical coordinates. In a first subsection, however, we discuss the Poisson structure of Hamiltonian systems in canonical form.

VII.2.1 Canonical Poisson Structure

... quelques remarques sur la plus profonde découverte de M. Poisson, mais qui, je crois, n'a pas été bien comprise ni par Lagrange, ni par les nombreux géomètres qui l'ont citée, ni par son auteur lui-même.

(C.G.J. Jacobi 1840, p. 350)

The derivative of a function $F(p, q)$ along the flow of a Hamiltonian system

$$\dot{p} = -\frac{\partial H}{\partial q}(p, q), \quad \dot{q} = \frac{\partial H}{\partial p}(p, q), \quad (2.1)$$

is given by (Lie derivative, see (III.5.3))

$$\frac{d}{dt}F(p(t), q(t)) = \sum_{i=1}^d \left(\frac{\partial F}{\partial p_i} \dot{p}_i + \frac{\partial F}{\partial q_i} \dot{q}_i \right) = \sum_{i=1}^d \left(\frac{\partial F}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial F}{\partial p_i} \frac{\partial H}{\partial q_i} \right). \quad (2.2)$$

This remarkably symmetric structure motivates the following definition.

Definition 2.1. The (canonical) *Poisson bracket* of two smooth functions $F(p, q)$ and $G(p, q)$ is the function

$$\{F, G\} = \sum_{i=1}^d \left(\frac{\partial F}{\partial q_i} \frac{\partial G}{\partial p_i} - \frac{\partial F}{\partial p_i} \frac{\partial G}{\partial q_i} \right), \quad (2.3)$$

or in vector notation $\{F, G\}(y) = \nabla F(y)^T J^{-1} \nabla G(y)$, where $y = (p, q)$ and J is the matrix of (VI.2.3).

This Poisson bracket is bilinear, skew-symmetric ($\{F, G\} = -\{G, F\}$), it satisfies the *Jacobi identity* (Jacobi 1862, *Werke* 5, p. 46)

$$\{\{F, G\}, H\} + \{\{G, H\}, F\} + \{\{H, F\}, G\} = 0 \quad (2.4)$$

(notice the cyclic permutations among F, G, H), and *Leibniz'* rule

$$\{F \cdot G, H\} = F \cdot \{G, H\} + G \cdot \{F, H\}. \quad (2.5)$$

These formulas are obtained in a straightforward manner from standard rules of calculus (see also Exercise 1).

With this notation, the Lie derivative (2.2) becomes

$$\frac{d}{dt}F(y(t)) = \{F, H\}(y(t)). \quad (2.6)$$

It follows that a function $I(p, q)$ is a first integral of (2.1) if and only if

$$\{I, H\} = 0.$$

If we take $F(y) = y_i$, the mapping that selects the i th component of y , we see that the Hamiltonian system (2.1) or (VI.2.5), $\dot{y} = J^{-1} \nabla H(y)$, can be written as

$$\dot{y}_i = \{y_i, H\}, \quad i = 1, \dots, 2d. \quad (2.7)$$

Poisson's Discovery. At the beginning of the 19th century, the hope of being able to integrate a given system of differential equations by analytic formulas faded more and more, and the energy of researchers went to the construction of, at least, first integrals. In this enthusiasm, Jacobi declared the subsequent result to be “Poisson's deepest discovery” (see citation) and his own identity, developed for its proof, a “gravissimum Theorema”.

Theorem 2.2 (Poisson 1809). *If I_1 and I_2 are first integrals, then their Poisson bracket $\{I_1, I_2\}$ is again a first integral.*

Proof. This follows at once from the Jacobi identity with $F = I_1$ and $G = I_2$. \square



Siméon Denis Poisson¹

VII.2.2 General Poisson Structures

... the general concept of a Poisson manifold should be credited to Sophus Lie in his treatise on transformation groups ...

(J.E. Marsden & T.S. Ratiu 1999)

We now come to the announced generalization of Definition 2.1 of the canonical Poisson bracket, invented by Lie (1888). Indeed, many proofs of properties of Hamiltonian systems rely uniquely on the bilinearity, the skew-symmetry and the Jacobi identity of the Poisson bracket, but not on the special structure of (2.3). So the idea is, more generally, to start with a smooth matrix-valued function $B(y) = (b_{ij}(y))$ and to set

$$\{F, G\}(y) = \sum_{i,j=1}^n \frac{\partial F(y)}{\partial y_i} b_{ij}(y) \frac{\partial G(y)}{\partial y_j} \quad (2.8)$$

(or more compactly $\{F, G\}(y) = \nabla F(y)^T B(y) \nabla G(y)$).

Lemma 2.3. *The bracket defined in (2.8) is bilinear, skew-symmetric and satisfies Leibniz' rule (2.5) as well as the Jacobi identity (2.4) if and only if*

$$b_{ij}(y) = -b_{ji}(y) \quad \text{for all } i, j \quad (2.9)$$

and for all i, j, k (notice the cyclic permutations among i, j, k)

$$\sum_{l=1}^n \left(\frac{\partial b_{ij}(y)}{\partial y_l} b_{lk}(y) + \frac{\partial b_{jk}(y)}{\partial y_l} b_{li}(y) + \frac{\partial b_{ki}(y)}{\partial y_l} b_{lj}(y) \right) = 0. \quad (2.10)$$

¹ Siméon Denis Poisson, born: 21 June 1781 in Pithiviers (France), died: 25 April 1840 in Sceaux (near Paris).

Proof. The main observation is that condition (2.10) is the Jacobi identity for the special choice of functions $F = y_i$, $G = y_j$, $H = y_k$ because of

$$\{y_i, y_j\} = b_{ij}(y). \quad (2.11)$$

If equation (2.4) is developed for the bracket (2.8), one obtains terms containing second order partial derivatives – these cancel due to the symmetry of the Jacobi identity – and terms containing first order partial derivatives; for the latter we may assume F, G, H to be linear combinations of y_i, y_j, y_k , so we are back to (2.10). The details of this proof are left as an exercise (see Exercise 1). \square

Definition 2.4. If the matrix $B(y)$ satisfies the properties of Lemma 2.3, formula (2.8) is said to represent a (general) *Poisson bracket*. The corresponding differential system

$$\dot{y} = B(y)\nabla H(y), \quad (2.12)$$

is a *Poisson system*. We continue to call H a Hamiltonian.

The system (2.12) can again be written in the bracket formulation (2.7). The formula (2.6) for the Lie derivative remains also valid, as is seen immediately from the chain rule and the definition of the Poisson bracket. Choosing $F = H$, this shows in particular that the Hamiltonian H is a first integral for general Poisson systems.

Definition 2.5. A function $C(y)$ is called a *Casimir function* of the Poisson system (2.12), if

$$\nabla C(y)^T B(y) = 0 \quad \text{for all } y.$$

A Casimir function is a first integral of every Poisson system with structure matrix $B(y)$, whatever the Hamiltonian $H(y)$ is.

Example 2.6. The *Lotka–Volterra* equations of Sect. I.1.1 can be written as

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} 0 & uv \\ -uv & 0 \end{pmatrix} \nabla H(u, v), \quad (2.13)$$

where $H(u, v) = u - \ln u + v - 2 \ln v$ is the invariant (I.1.4). This is of the form (2.12) with a matrix that is skew-symmetric and satisfies the identity (2.10).

Higher dimensional Lotka–Volterra systems can also have a Poisson structure (see, e.g., Perelomov (1995) and Suris (1999)). For example, the system

$$\dot{y}_1 = y_1(y_2 + y_3), \quad \dot{y}_2 = y_2(y_1 - y_3 + 1), \quad \dot{y}_3 = y_3(y_1 + y_2 + 1)$$

can be written as

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & y_1 y_2 & y_1 y_3 \\ -y_1 y_2 & 0 & -y_2 y_3 \\ -y_1 y_3 & y_2 y_3 & 0 \end{pmatrix} \nabla H(y) \quad (2.14)$$

with $H(y) = -y_1 + y_2 + y_3 + \ln y_2 - \ln y_3$. Again one can check by direct computation that (2.10) is satisfied.

In contrast to the structure matrix J^{-1} of Hamiltonian systems in canonical form, the matrix $B(y)$ of (2.12) need not be invertible. All odd-dimensional skew-symmetric matrices are singular, and so is the matrix $B(y)$ of (2.14). In this case, the vector $v(y) = (-1/y_1, -1/y_2, 1/y_3)^T$ satisfies $v(y)^T B(y) = 0$. Since $v(y) = \nabla C(y)$ with $C(y) = -\ln y_1 - \ln y_2 + \ln y_3$, the function $C(y)$ is a Casimir function.

VII.2.3 Hamiltonian Systems on Symplectic Submanifolds

An important motivation for studying Poisson systems is given by Hamiltonian problems expressed in non-canonical coordinates.

Example 2.7 (Constrained Mechanical Systems). Consider the system (1.9) written as the differential equation

$$\dot{x} = J^{-1} \left(\nabla H(x) + \sum_{i=1}^m \lambda_i(x) \nabla g_i(x) \right) \quad (2.15)$$

on the manifold $\mathcal{M} = \{x; c(x) = 0\}$ with $c(x) = (g(q), G(q)H_p(p, q))^T$ and $x = (p, q)^T$ (see (1.14)). As in the proof of Theorem 1.2, $\lambda_i(x)$ and $g_i(x)$ are the components of $\lambda(x)$ and $g(x)$, and $\lambda(x)$ is the function obtained from (1.12). We use $y \in \mathbb{R}^{2(d-m)}$ as local coordinates of the manifold \mathcal{M} via the transformation

$$x = \chi(y).$$

In these coordinates, the differential equation (2.15) becomes, with $X(y) = \chi'(y)$,

$$X(y) \dot{y} = J^{-1} \left(\nabla H(\chi(y)) + \sum_{i=1}^m \lambda_i(\chi(y)) \nabla g_i(\chi(y)) \right).$$

We multiply this equation from the left with $X(y)^T J$ and note that the columns of $X(y)$, which are tangent vectors, are orthogonal to the gradients ∇g_i of the constraints. This yields

$$X(y)^T J X(y) \dot{y} = X(y)^T \nabla H(\chi(y)).$$

By assumption (1.13) the matrix $X(y)^T J X(y)$ is invertible. This is seen as follows: $X(y)^T J X(y)v = 0$ implies $JX(y)v = c'(x)^T w$ for some w ($x = \chi(y)$). By $c(\chi(y)) = 0$ and $c'(x)X(y) = 0$ we get $c'(x)J^{-1}c'(x)^T w = 0$. It then follows from the structure of $c'(x)$ and from (1.13) that $w = 0$ and hence also $v = 0$.

With $B(y) = (X(y)^T J X(y))^{-1}$ and $K(y) = H(\chi(y))$, the above equation for \dot{y} thus becomes the Poisson system $\dot{y} = B(y) \nabla K(y)$. The matrix $B(y)$ is skew-symmetric and satisfies (2.10), see Theorem 2.8 below or Exercise 11.

More generally, consider a *symplectic submanifold* \mathcal{M} of \mathbb{R}^{2d} , that is, a manifold for which the symplectic two-form²

$$\omega_x(\xi_1, \xi_2) = (J\xi_1, \xi_2) \quad \text{for } \xi_1, \xi_2 \in T_x\mathcal{M} \quad (2.16)$$

(with (\cdot, \cdot) denoting the Euclidean inner product on \mathbb{R}^{2d}) is *non-degenerate* for every $x \in \mathcal{M}$: for ξ_1 in the tangent space $T_x\mathcal{M}$,

$$\omega_x(\xi_1, \xi_2) = 0 \quad \text{for all } \xi_2 \in T_x\mathcal{M} \quad \text{implies} \quad \xi_1 = 0.$$

In local coordinates $x = \chi(y)$, this condition is equivalent to the invertibility of the matrix $X(y)^T J X(y)$ with $X(y) = \chi'(y)$, since every tangent vector at $x = \chi(y)$ is of the form $\xi = X(y)\eta$ and $X(y)$ has linearly independent columns. A manifold defined by constraints, $\mathcal{M} = \{x \in \mathbb{R}^{2d} \mid c(x) = 0\}$, is symplectic if the matrix $c'(x)J^{-1}c'(x)^T$ is invertible for every $x \in \mathcal{M}$ (see the argument at the end of the previous example). This condition can be restated as saying that the matrix $(\{c_i, c_j\}(x))$ of canonical Poisson brackets of the constraint functions is invertible.

We consider the reduction of the Hamiltonian system to the symplectic submanifold \mathcal{M} , which determines solution curves $t \mapsto x(t) \in \mathcal{M}$ by the equations

$$(J\dot{x} - \nabla H(x), \xi) = 0 \quad \text{for all } \xi \in T_x\mathcal{M}. \quad (2.17)$$

With the interpretation $(\nabla H(x), \xi) = H'(x)\xi = \frac{d}{dt}|_{t=0} H(\gamma(t))$ as a directional derivative along a path $\gamma(t) \in \mathcal{M}$ with $\gamma(0) = x$ and $\dot{\gamma}(0) = \xi$, it is sufficient that the Hamiltonian H is defined and differentiable on the manifold \mathcal{M} . Equation (2.17) can also be expressed as

$$\omega_x(\dot{x}, \xi) = H'(x)\xi \quad \text{for all } \xi \in T_x\mathcal{M}, \quad (2.18)$$

a formulation that is susceptible to further generalization; cf. Marsden & Ratiu (1999), Chap. 5.4, and Exercise 2. Choosing $\xi = \dot{x}$ we obtain $0 = H'(x)\dot{x} = \frac{d}{dt} H(x(t))$, and hence the Hamiltonian is conserved along solutions.

Note that for \mathcal{M} of Example 2.7, the formulation (2.17) is equivalent to the equations of motion (2.15) of the constrained mechanical system. It corresponds to *d'Alembert's principle of virtual variations* in constrained mechanics; see Arnold (1989), p. 92. In quantum mechanics the Hamiltonian reduction (2.17) to a manifold (in that case, a submanifold of the Hilbert space $L^2(\mathbb{R}^N, \mathbb{R}^2)$ instead of \mathbb{R}^{2d}) is known as the *Dirac–Frenkel time-dependent variational principle* and is the basic tool for deriving reduced models of the many-body Schrödinger equation; see Sect. VII.6 for an example. From a numerical analysis viewpoint, (2.17) can also be viewed as a Galerkin method on the solution-dependent tangent space $T_x\mathcal{M}$.

In terms of the *symplectic projection* $P(x) : \mathbb{R}^{2d} \rightarrow T_x\mathcal{M}$ for $x \in \mathcal{M}$, defined by determining $P(x)v \in T_x\mathcal{M}$ for $v \in \mathbb{R}^{2d}$ from the condition

$$(JP(x)v, \xi) = (Jv, \xi) \quad \text{for all } \xi \in T_x\mathcal{M}, \quad (2.19)$$

² Notice that this two-form is the negative of that introduced in Sect. VI.2. This slight inconsistency makes the subsequent formulas nicer.

formula (2.17) can be reformulated as the differential equation on \mathcal{M} ,

$$\dot{x} = P(x)J^{-1}\nabla H(x). \quad (2.20)$$

In coordinates $x = \chi(y)$, and again with $X(y) = \chi'(y)$, formula (2.17) becomes

$$X(y)^T \left(JX(y)\dot{y} - \nabla H(\chi(y)) \right) = 0,$$

and with

$$B(y) = \left(X(y)^T JX(y) \right)^{-1} \quad \text{and} \quad K(y) = H(\chi(y)), \quad (2.21)$$

we obtain the differential equation

$$\dot{y} = B(y)\nabla K(y). \quad (2.22)$$

Theorem 2.8. *For a Hamiltonian system (2.17) on a symplectic submanifold \mathcal{M} , the equivalent differential equation in local coordinates, (2.22) with (2.21), is a Poisson system.*

Proof. In coordinates, the symplectic projection is given by

$$P(x) = X(y)B(y)X(y)^T J \quad \text{for } x = \chi(y) \in \mathcal{M},$$

since for every tangent vector $\xi = X(y)\eta$ we have by (2.21),

$$(JXB X^T Jv, X\eta) = (X^T JXB X^T Jv, \eta) = (X^T Jv, \eta) = (Jv, X\eta).$$

From the decomposition $\mathbb{R}^{2d} = P(x)\mathbb{R}^{2d} \oplus (I - P(x))\mathbb{R}^{2d}$ we obtain, by the implicit function theorem, a corresponding splitting in a neighbourhood of the manifold \mathcal{M} in \mathbb{R}^{2d} ,

$$v = x + w \quad \text{with } x \in \mathcal{M}, P(x)w = 0.$$

This permits us to extend smooth functions $F(y)$ to a neighbourhood of \mathcal{M} by setting

$$\widehat{F}(v) = F(y) \quad \text{for } v = x + w \text{ with } x = \chi(y), P(x)w = 0.$$

We then have for the derivative $\widehat{F}'(x) = \widehat{F}'(x)P(x)$ for $x \in \mathcal{M}$ and hence for its transpose, the gradient, $\nabla \widehat{F}(x) = P(x)^T \nabla \widehat{F}(x)$. Moreover, by the chain rule we have $\nabla F(y) = X(y)^T \nabla \widehat{F}(x)$ for $x = \chi(y)$. For the canonical bracket this gives, at $x = \chi(y)$,

$$\begin{aligned} \{\widehat{F}, \widehat{G}\}_{\text{can}}(x) &= \nabla \widehat{F}(x)^T P(x) J^{-1} P(x)^T \nabla \widehat{G}(x) \\ &= \nabla F(y)^T B(y) \nabla G(y) = \{F, G\}(y), \end{aligned}$$

and hence the required properties of the bracket defined by $B(y)$ follow from the corresponding properties of the canonical bracket. \square

VII.3 The Darboux–Lie Theorem

Theorem 2.8 also shows that a Hamiltonian system without constraints becomes a Poisson system in non-canonical coordinates. Interestingly, a converse also holds: every Poisson system can locally be written in canonical Hamiltonian form after a suitable change of coordinates. This result is a special case of the *Darboux–Lie Theorem*. Its proof was the result of several important papers: Jacobi’s theory of simultaneous linear partial differential equations (Jacobi 1862), the works by Clebsch (1866) and Darboux (1882) on Pfaffian systems, and, finally, the paper of Lie (1888). We shall now retrace this development. Our first tool is a result on the commutativity of Poisson flows.

VII.3.1 Commutativity of Poisson Flows and Lie Brackets

The elegant formula (2.6) for the Lie derivative is valid for general Poisson systems with the vector field $f(y) = B(y)\nabla H(y)$ of (2.12). Acting on a function $F : \mathbb{R}^n \rightarrow \mathbb{R}$, the Lie operator (III.5.2) becomes

$$DF = \nabla F^T f = \nabla F^T B(y)\nabla H = \{F, H\} \quad (3.1)$$

and is again the Poisson bracket. This observation is the key for the following lemma, which shows an interesting connection between the Lie bracket and the Poisson bracket.

Lemma 3.1. *Let two smooth Hamiltonians $H^{[1]}(y)$ and $H^{[2]}(y)$ be given.*

$$\begin{aligned} \text{If } D_1 & \text{ is the Lie operator of } B(y)\nabla H^{[1]} \\ \text{and } D_2 & \text{ is the Lie operator of } B(y)\nabla H^{[2]}, \\ \text{then } [D_1, D_2] & \text{ is the Lie operator of } B(y)\nabla\{H^{[2]}, H^{[1]}\} \end{aligned} \quad (3.2)$$

(notice, once again, that the indices 1 and 2 have been reversed).

Proof. After some clever permutations, the Jacobi identity (2.4) can be written as

$$\{\{F, H^{[2]}\}, H^{[1]}\} - \{\{F, H^{[1]}\}, H^{[2]}\} = \{F, \{H^{[2]}, H^{[1]}\}\}. \quad (3.3)$$

By (3.1) this is nothing other than $D_1 D_2 F - D_2 D_1 F = [D_1, D_2]F$. \square

Lemma 3.2. *Consider two smooth Hamiltonians $H^{[1]}(y)$ and $H^{[2]}(y)$ on an open connected set U , with D_1 and D_2 the corresponding Lie operators and $\varphi_s^{[1]}(y)$ and $\varphi_t^{[2]}(y)$ the corresponding flows. Then, if the matrix $B(y)$ is invertible, the following are equivalent in U :*

- (i) $\{H^{[1]}, H^{[2]}\} = \text{Const}$;
- (ii) $[D_1, D_2] = 0$;
- (iii) $\varphi_t^{[2]} \circ \varphi_s^{[1]} = \varphi_s^{[1]} \circ \varphi_t^{[2]}$.

The conclusions “(i) \Rightarrow (ii) \Leftrightarrow (iii)” also hold for a non-invertible $B(y)$.

Proof. This is obtained by combining Lemma III.5.4 and Lemma 3.1. We need the invertibility of $B(y)$ to conclude that $\{H^{[1]}, H^{[2]}\} = \text{Const}$ follows from $B(y)\nabla\{H^{[1]}, H^{[2]}\} = 0$. \square

VII.3.2 Simultaneous Linear Partial Differential Equations

If two functions $F(y)$ and $G(y)$ are given, formula (2.8) determines a function $h(y) = \{F, G\}(y)$ by differentiation. We now ask the *inverse* question: Given functions $G(y)$ and $h(y)$, can we find a function $F(y)$ such that $\{F, G\}(y) = h(y)$? This problem represents a first order linear partial differential equation for F . So we are led to the following problem, which we first discuss in two dimensions.

One Equation. Given functions $a(y_1, y_2)$, $b(y_1, y_2)$, $h(y_1, y_2)$, find all solutions $F(y_1, y_2)$ satisfying

$$a(y_1, y_2) \frac{\partial F}{\partial y_1} + b(y_1, y_2) \frac{\partial F}{\partial y_2} = h(y_1, y_2). \quad (3.4)$$

This equation is, for any point (y_1, y_2) , a linear relation between the partial derivatives of F , but does not determine them individually. There is *one* direction, however, where the derivative is uniquely determined, namely that of the vector $\mathbf{n} = (a(y_1, y_2), b(y_1, y_2))$, since the left-hand side of equation (3.4) is the directional derivative $\frac{\partial F}{\partial \mathbf{n}}$. The lines, which everywhere respect this direction, are called *characteristic lines* (see left picture of Fig. 3.1). If we parametrize them with a parameter t , we can compute $y_1(t)$, $y_2(t)$ as well as $F(t) = F(y_1(t), y_2(t))$ as solutions of the following ordinary differential equations

$$\dot{y}_1 = a(y_1, y_2), \quad \dot{y}_2 = b(y_1, y_2), \quad \dot{F} = h(y_1, y_2). \quad (3.5)$$

The *initial values* $(y_1(0), y_2(0))$ can be chosen on an arbitrary curve γ (which must be transversal to the characteristic lines) and the values $F|_\gamma$ can be arbitrarily prescribed. The solution $F(y_1, y_2)$ of (3.4) is then created by the curves (3.5) wherever the characteristic lines go (right picture of Fig. 3.1).

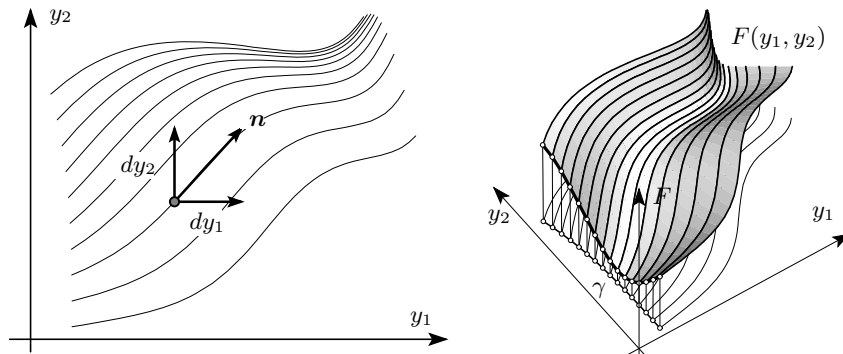


Fig. 3.1. Characteristic lines and solution of a first order linear partial differential equation

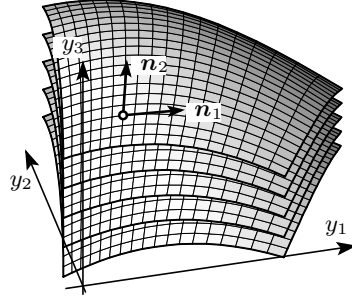


Fig. 3.2. Characteristic surfaces of two first order linear partial differential equations

For one equation in n dimensions, the initial values $(y_1(0), \dots, y_n(0))$ can be freely chosen on a manifold of dimension $n - 1$ (e.g., the subspace orthogonal to the characteristic line passing through a given point), and F can be arbitrarily prescribed on this manifold. This guarantees the existence of $n - 1$ independent solutions in the neighbourhood of a given point. Here, independent means that the gradients of these functions are linearly independent.

Two Simultaneous Equations. Two simultaneous equations of dimension two are trivial. We therefore suppose $y = (y_1, y_2, y_3)$ and two equations of the form

$$\begin{aligned} a_1^{[1]}(y) \frac{\partial F}{\partial y_1} + a_2^{[1]}(y) \frac{\partial F}{\partial y_2} + a_3^{[1]}(y) \frac{\partial F}{\partial y_3} &= h_1(y), \\ a_1^{[2]}(y) \frac{\partial F}{\partial y_1} + a_2^{[2]}(y) \frac{\partial F}{\partial y_2} + a_3^{[2]}(y) \frac{\partial F}{\partial y_3} &= h_2(y) \end{aligned} \quad (3.6)$$

for an unknown function $F(y_1, y_2, y_3)$. This system can also be written as $D_1 F = h_1$, $D_2 F = h_2$, where D_i denotes the Lie operator corresponding to the vector field $a^{[i]}(y)$. Here, we have *two* directional derivatives prescribed, namely $\frac{\partial F}{\partial \mathbf{n}_1}$ and $\frac{\partial F}{\partial \mathbf{n}_2}$ where $\mathbf{n}_i = a^{[i]}(y)$ (see Fig. 3.2). Therefore, we will have to follow both directions and, instead of (3.5), we will have *two* sets of ordinary differential equations

$$\begin{aligned} \dot{y}_1 &= a_1^{[1]}(y), & \dot{y}_2 &= a_2^{[1]}(y), & \dot{y}_3 &= a_3^{[1]}(y), & \dot{F} &= h_1(y) \\ \dot{y}_1 &= a_1^{[2]}(y), & \dot{y}_2 &= a_2^{[2]}(y), & \dot{y}_3 &= a_3^{[2]}(y), & \dot{F} &= h_2(y). \end{aligned} \quad (3.7)$$

If we prescribe F on a curve that is orthogonal to \mathbf{n}_1 and \mathbf{n}_2 , and if we follow the solutions of (3.7), we obtain the function F on two 2-dimensional surfaces S_1 and S_2 containing the prescribed curve. Continuing from S_1 along the second flow and from S_2 along the first flow, we may be led to the same point, but nothing guarantees that the obtained values for F are identical. To get a well-defined F , additional assumptions on the differential operators and on the inhomogeneities have to be made.

The following theorem, which is due to Jacobi (1862), has been extended by Clebsch (1866), who created the theory of *complete systems* (“vollständige

Systeme"). These papers contained long analytic calculations with myriades of formulas. The wonderful geometric insight is mainly due to Sophus Lie.

Theorem 3.3. *Let D_1, \dots, D_m be m ($m < n$) linear differential operators in \mathbb{R}^n corresponding to vector fields $a^{[1]}(y), \dots, a^{[m]}(y)$ and suppose that these vectors are linearly independent for $y = y_0$. If*

$$[D_i, D_j] = 0 \quad \text{for all } i, j, \quad (3.8)$$

then the homogeneous system

$$D_i F = 0 \quad \text{for } i = 1, \dots, m$$

possesses (in a neighbourhood of y_0) $n - m$ solutions for which the gradients $\nabla F(y_0)$ are linearly independent.

Furthermore, the inhomogeneous system of partial differential equations

$$D_i F = h_i \quad \text{for } i = 1, \dots, m$$

possesses a particular solution in a neighbourhood of y_0 , if and only if in addition to (3.8) the functions $h_1(y), \dots, h_m(y)$ satisfy the integrability conditions

$$D_i h_j = D_j h_i \quad \text{for all } i, j. \quad (3.9)$$

Proof. (a) Let V denote the space of vectors in \mathbb{R}^n that are orthogonal to $a^{[1]}(y_0), \dots, a^{[m]}(y_0)$, and consider the $(n - m)$ -dimensional manifold $\mathcal{M} = y_0 + V$. We then extend an arbitrary smooth function $F : \mathcal{M} \rightarrow \mathbb{R}$ to a neighbourhood of y_0 by

$$F(\varphi_{t_m}^{[m]} \circ \dots \circ \varphi_{t_1}^{[1]}(y_0 + v)) = F(y_0 + v). \quad (3.10)$$

Notice that $(t_1, \dots, t_m, v) \mapsto y = \varphi_{t_m}^{[m]} \circ \dots \circ \varphi_{t_1}^{[1]}(y_0 + v)$ defines a local diffeomorphism between neighbourhoods of 0 and y_0 . Since the application of the operator D_m to (3.10) corresponds to a differentiation with respect to t_m and the expression $F(\varphi_{t_m}^{[m]} \circ \dots \circ \varphi_{t_1}^{[1]}(y_0 + v))$ is independent of t_m by (3.10), we get $D_m F(y) = 0$. To prove $D_i F(y) = 0$ for $i < m$, we first have to change the order of the flows $\varphi_{t_j}^{[j]}$ in (3.10), which is permitted by Lemma III.5.4 and assumption (3.8), so that $\varphi_{t_i}^{[i]}$ is in the left-most position.

(b) The necessity of (3.9) follows immediately from $D_i h_j = D_i D_j F = D_j D_i F = D_j h_i$. For given h_i satisfying (3.9) we define $F(y)$ in a neighbourhood of y_0 (i.e., for small t_1, \dots, t_m and small v) by

$$\begin{aligned} F(\varphi_{t_m}^{[m]} \circ \dots \circ \varphi_{t_1}^{[1]}(y_0 + v)) &= \int_0^{t_1} h_1(\varphi_t^{[1]}(y_0 + v)) dt \\ &+ \dots + \int_0^{t_m} h_m(\varphi_t^{[m]} \circ \varphi_{t_{m-1}}^{[m-1]} \circ \dots \circ \varphi_{t_1}^{[1]}(y_0 + v)) dt, \end{aligned}$$

and we prove that it is a solution of the system $D_i F = h_i$ for $i = 1, \dots, m$. Since only the last integral depends on t_m , we immediately get by differentiation with respect to t_m that $D_m F = h_m$. For the computation of $D_i F$ we differentiate with respect to t_i . The first $i - 1$ integrals are independent of t_i . The derivative of the i th integral gives $h_i(\varphi_{t_i}^{[i]} \circ \dots \circ \varphi_{t_1}^{[1]}(y_0 + v))$, and the derivative of the remaining integrals gives

$$\begin{aligned} \int_0^{t_j} D_i h_j(\varphi_t^{[j]} \circ \dots \circ \varphi_{t_1}^{[1]}(y_0 + v)) dt &= \int_0^{t_j} D_j h_i(\varphi_t^{[j]} \circ \dots \circ \varphi_{t_1}^{[1]}(y_0 + v)) dt \\ &= h_i(\varphi_{t_j}^{[j]} \circ \dots \circ \varphi_{t_1}^{[1]}(y_0 + v)) - h_i(\varphi_{t_{j-1}}^{[j-1]} \circ \dots \circ \varphi_{t_1}^{[1]}(y_0 + v)) \end{aligned}$$

for $j = i + 1, \dots, m$. Summing up, this proves $D_i F = h_i$. \square

VII.3.3 Coordinate Changes and the Darboux–Lie Theorem

The emphasis here is to simplify a given Poisson structure as much as possible by a coordinate transformation. We change from coordinates y_1, \dots, y_n to $\tilde{y}_1(y), \dots, \tilde{y}_n(y)$ with continuously differentiable functions and an invertible Jacobian $A(y) = \partial \tilde{y} / \partial y$,

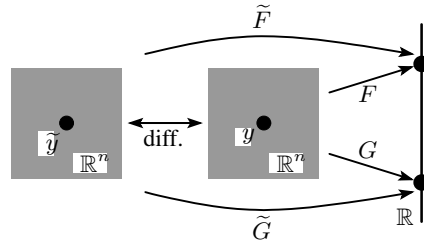


Fig. 3.3. New coordinates in a Poisson system



Jean Gaston Darboux³

and we denote $\tilde{F}(\tilde{y}) := F(y)$ and $\tilde{G}(\tilde{y}) := G(y)$ (see Fig. 3.3). The Poisson structure as well as the Poisson flow on one space will become another Poisson structure and flow on the other space by simply applying the chain rule:

$$\sum_{i,j} \frac{\partial F(y)}{\partial y_i} b_{ij}(y) \frac{\partial G(y)}{\partial y_j} = \sum_{i,j,k,l} \frac{\partial \tilde{F}(\tilde{y})}{\partial \tilde{y}_k} \frac{\partial \tilde{y}_k}{\partial y_i} b_{ij}(y(\tilde{y})) \frac{\partial \tilde{y}_l}{\partial y_j} \frac{\partial \tilde{G}(\tilde{y})}{\partial \tilde{y}_l}. \quad (3.11)$$

This is another Poisson structure with

$$\tilde{b}_{kl} = \{\tilde{y}_k, \tilde{y}_l\} \quad \text{or} \quad \tilde{B}(\tilde{y}) = A(y)B(y)A(y)^T. \quad (3.12)$$

³ Jean Gaston Darboux, born: 14 August 1842 in Nîmes (France), died: 23 February 1917 in Paris.

The same structure matrix is obtained if the Poisson system (2.12) is written in these new coordinates (Exercise 5).

Since A is invertible, the structure matrices B and \tilde{B} have the same rank. We now want to obtain the simplest possible form for \tilde{B} .

Theorem 3.4 (Darboux 1882, Lie 1888). *Suppose that the matrix $B(y)$ defines a Poisson bracket and is of constant rank $n - q = 2m$ in a neighbourhood of $y_0 \in \mathbb{R}^n$. Then, there exist functions $P_1(y), \dots, P_m(y)$, $Q_1(y), \dots, Q_m(y)$, and $C_1(y), \dots, C_q(y)$ satisfying*

$$\begin{aligned} \{P_i, P_j\} &= 0 & \{P_i, Q_j\} &= -\delta_{ij} & \{P_i, C_l\} &= 0 \\ \{Q_i, P_j\} &= \delta_{ij} & \{Q_i, Q_j\} &= 0 & \{Q_i, C_l\} &= 0 \\ \{C_k, P_j\} &= 0 & \{C_k, Q_j\} &= 0 & \{C_k, C_l\} &= 0 \end{aligned} \quad (3.13)$$

on a neighbourhood of y_0 . The gradients of P_i, Q_i, C_k are linearly independent, so that $y \mapsto (P_i(y), Q_i(y), C_k(y))$ constitutes a local change of coordinates to canonical form.

The functions $C_1(y), \dots, C_q(y)$ are called *distinguished functions* (ausgezeichnete Funktionen) by Lie.

Proof. We follow Lie's original proof. Similar ideas, and the same notation, are also present in Darboux's paper. The proof proceeds in several steps, satisfying the conditions of (3.13), from one line to the next, by solving systems of linear partial differential equations.

(a) If all $b_{ij}(y_0) = 0$, the constant rank assumption implies $b_{ij}(y) = 0$ in a neighbourhood of y_0 . We thus have $m = 0$ and all coordinates $C_i(y) = y_i$ are Casimirs.

(b) If there exist i, j with $b_{ij}(y_0) \neq 0$, we set $Q_1(y) = y_i$ and we determine $P_1(y)$ as the solution of the linear partial differential equation

$$\{Q_1, P_1\} = 1. \quad (3.14)$$

Because of $b_{ij}(y_0) \neq 0$ the assumption of Theorem 3.3 is satisfied and this yields the existence of P_1 . We next consider the homogeneous system

$$\{Q_1, F\} = 0 \quad \text{and} \quad \{P_1, F\} = 0 \quad (3.15)$$

of partial differential equations. By Lemma 3.2 and (3.14) the Lie operators corresponding to Q_1 and P_1 commute, so that by Theorem 3.3 the system (3.15) has $n - 2$ independent solutions F_3, \dots, F_n . Their gradients together with those of Q_1 and P_1 form a basis of \mathbb{R}^n . We therefore can change coordinates from y_1, \dots, y_n to $Q_1, P_1, F_3, \dots, F_n$ (mapping y_0 to \tilde{y}_0). In these coordinates the first two rows and the first two columns of the structure matrix $\tilde{B}(\tilde{y})$ have the required form.

(c) If $\tilde{b}_{ij}(\tilde{y}_0) = 0$ for all $i, j \geq 3$, we have $m = 1$ (similar to step (a)) and the coordinates F_3, \dots, F_n are Casimirs.

(d) If there exist $i \geq 3$ and $j \geq 3$ with $\tilde{b}_{ij}(\tilde{y}_0) \neq 0$, we set $Q_2 = F_i$ and we determine P_2 from the inhomogeneous system

$$\{Q_1, P_2\} = 0, \quad \{P_1, P_2\} = 0, \quad \{Q_2, P_2\} = 1.$$

The inhomogeneities satisfy (3.9), and the Lie operators corresponding to Q_1, P_1, Q_2 commute (by Lemma 3.2). Theorem 3.3 proves the existence of such a P_2 . We then consider the homogeneous system

$$\{Q_1, F\} = 0, \quad \{P_1, F\} = 0, \quad \{Q_2, F\} = 0, \quad \{P_2, F\} = 0$$

and apply once more Theorem 3.3. We get $n - 4$ independent solutions, which we denote again F_5, \dots, F_n . As in part (b) of the proof we get new coordinates $Q_1, P_1, Q_2, P_2, F_5, \dots, F_n$, for which the first *four* rows and columns of the structure matrix are canonical.

(e) The proof now continues by repeating steps (c) and (d) until the structure matrix has the desired form. \square

Corollary 3.5 (Casimir Functions). *In the situation of Theorem 3.4 the functions $C_1(y), \dots, C_q(y)$ satisfy*

$$\{C_i, H\} = 0 \quad \text{for all smooth } H. \quad (3.16)$$

Proof. Theorem 3.4 states that $\nabla C_i(y)^T B(y) \nabla H(y) = 0$, when $H(y)$ is one of the functions $P_j(y), Q_j(y)$ or $C_j(y)$. However, the gradients of these functions form a basis of \mathbb{R}^n . Consequently, $\nabla C_i(y)^T B(y) = 0$ and (3.16) is satisfied for all differentiable functions $H(y)$. \square

This property implies that all Casimir functions are first integrals of (2.12) whatever $H(y)$ is. Consequently, (2.12) is (close to y_0) a differential equation on the manifold

$$\mathcal{M} = \{y \in U \mid C_i(y) = \text{Const}_i, i = 1, \dots, m\}. \quad (3.17)$$

Corollary 3.6 (Transformation to Canonical Form). *Denote the transformation of Theorem 3.4 by $z = \vartheta(y) = (P_i(y), Q_i(y), C_k(y))$. With this change of coordinates, the Poisson system $\dot{y} = B(y) \nabla H(y)$ becomes*

$$\dot{z} = B_0 \nabla K(z) \quad \text{with} \quad B_0 = \begin{pmatrix} J^{-1} & 0 \\ 0 & 0 \end{pmatrix}, \quad (3.18)$$

where $K(z) = H(y)$. Writing $z = (p, q, c)$, this system becomes

$$\dot{p} = -K_q(p, q, c), \quad \dot{q} = K_p(p, q, c), \quad \dot{c} = 0.$$

Proof. The transformed differential equation is

$$\dot{z} = \vartheta'(y) B(y) \vartheta'(y)^T \nabla K(z) \quad \text{with} \quad y = \vartheta^{-1}(z),$$

and Theorem 3.4 states that $\vartheta'(y) B(y) \vartheta'(y)^T = B_0$. \square

VII.4 Poisson Integrators

Before discussing geometric numerical integrators, we show that many important properties of Hamiltonian systems in canonical form remain valid for systems

$$\dot{y} = B(y)\nabla H(y), \quad (4.1)$$

where $B(y)$ represents a Poisson bracket.

VII.4.1 Poisson Maps and Symplectic Maps

We have already seen that the Hamiltonian $H(y)$ is a first integral of (4.1). We shall show here that the flow of (4.1) satisfies a property closely related to symplecticity.

Definition 4.1. A transformation $\varphi : U \rightarrow \mathbb{R}^n$ (where U is an open set in \mathbb{R}^n) is called a *Poisson map* with respect to the bracket (2.8), if its Jacobian matrix satisfies

$$\varphi'(y)B(y)\varphi'(y)^T = B(\varphi(y)). \quad (4.2)$$

An equivalent condition is that for all smooth real-valued functions F, G defined on $\varphi(U)$,

$$\{F \circ \varphi, G \circ \varphi\}(y) = \{F, G\}(\varphi(y)), \quad (4.3)$$

as is seen by the chain rule and choosing F, G as the coordinate functions. It is clear from this condition that the composition of Poisson maps is again a Poisson map. A comparison with (3.12) shows that Poisson maps leave the structure matrix invariant.

For the canonical symplectic structure, where $B(y) = J^{-1}$, condition (4.2) is equivalent to the symplecticity of the transformation $\varphi(y)$. This can be seen by taking the inverse of both sides of (4.2), and by multiplying the resulting equation with $\varphi'(y)$ from the right and with $\varphi'(y)^T$ from the left. Also in the situation of a Hamiltonian system (2.17) on a symplectic submanifold \mathcal{M} , where $B(y)$ is the structure matrix of the differential equation in coordinates y as in Theorem 2.8, condition (4.2) is equivalent to symplecticity in the sense of preserving the symplectic two-form (2.16) on the tangent space, as in (1.16):

Definition 4.2. A map $\psi : \mathcal{M} \rightarrow \mathcal{M}$ on a symplectic manifold \mathcal{M} is called *symplectic* if for every $x \in \mathcal{M}$,

$$\omega_{\psi(x)}(\psi'(x)\xi_1, \psi'(x)\xi_2) = \omega_x(\xi_1, \xi_2) \quad \text{for all } \xi_1, \xi_2 \in T_x\mathcal{M}. \quad (4.4)$$

A near-identity map $\psi : \mathcal{M} \rightarrow \mathcal{M}$ is symplectic if and only if the conjugate map φ in local coordinates $x = \chi(y)$, with $\varphi(y)$ given by $\psi(x) = \chi(\varphi(y))$ for $x = \chi(y)$, is a Poisson map for the structure matrix of (2.21), $B(y) = (X(y)^T J X(y))^{-1}$ with $X(y) = \chi'(y)$. This holds because $\psi'(x)\xi = X(\varphi(y))\varphi'(y)\eta$ for $x = \chi(y)$ and $\xi = X(y)\eta$, and because (4.2) is equivalent to $\varphi'(y)^T X(\varphi(y))^T J X(\varphi(y))\varphi'(y) = X(y)^T J X(y)$.

Theorem 4.3. *If $B(y)$ is the structure matrix of a Poisson bracket, then the flow $\varphi_t(y)$ of the differential equation (4.1) is a Poisson map.*

Proof. (a) For $B(y) = J^{-1}$ this is exactly the statement of Theorem VI.2.4 on the symplecticity of the flow of Hamiltonian systems. This result can be extended in a straightforward way to the matrix B_0 of (3.18).

(b) For the general case consider the change of coordinates $z = \vartheta(y)$ which transforms (4.1) to canonical form (Theorem 3.4), i.e., $\vartheta'(y)B(y)\vartheta'(y)^T = B_0$ and $\dot{z} = B_0 \nabla K(z)$ with $K(z) = H(y)$ (Corollary 3.6). Denoting the flows of (4.1) and $\dot{z} = B_0 \nabla K(z)$ by $\varphi_t(y)$ and $\psi_t(z)$, respectively, we have $\psi_t(\vartheta(y)) = \vartheta(\varphi_t(y))$ and by the chain rule $\psi'_t(\vartheta(y))\vartheta'(y) = \vartheta'(\varphi_t(y))\varphi'_t(y)$. Inserting this relation into $\psi'_t(z)B_0\psi'_t(z)^T = B_0$, which follows from (a), proves the statement.

A direct proof, avoiding the use of Theorem 3.4, is indicated in Exercise 6. \square

From Theorems 2.8 and 4.3 and the remark after Definition 4.2 we note the following.

Corollary 4.4. *The flow of a Hamiltonian system (2.17) on a symplectic submanifold is symplectic.*

The inverse of Theorem 4.3 is also true. It extends Theorem VI.2.6 from canonically symplectic transformations to Poisson maps.

Theorem 4.5. *Let $f(y)$ and $B(y)$ be continuously differentiable on an open set $U \subset \mathbb{R}^m$, and assume that $B(y)$ represents a Poisson bracket (Definition 2.4). Then, $\dot{y} = f(y)$ is locally of the form (4.1), if and only if*

- *its flow $\varphi_t(y)$ respects the Casimirs of $B(y)$, i.e., $C_i(\varphi_t(y)) = \text{Const}$, and*
- *its flow is a Poisson map for all $y \in U$ and for all sufficiently small t .*

Proof. The necessity follows from Corollary 3.5 and from Theorem 4.3. For the proof of sufficiency we apply the change of coordinates $(u, c) = \vartheta(y)$ of Theorem 3.4, which transforms $B(y)$ into canonical form (3.18). We write the differential equation $\dot{y} = f(y)$ in the new variables as

$$\dot{u} = g(u, c), \quad \dot{c} = h(u, c). \quad (4.5)$$

Our first assumption expresses the fact that the Casimirs, which are the components of c , are first integrals of this system. Consequently, we have $h(u, c) \equiv 0$. The second assumption implies that the flow of (4.5) is a Poisson map for B_0 of (3.18). Writing down explicitly the blocks of condition (4.2), we see that this is equivalent to the symplecticity of the mapping $u_0 \mapsto u(t, u_0, c_0)$, with c_0 as a parameter. From Theorem VI.2.6 we thus obtain the existence of a function $K(u, c)$ such that $g(u, c) = J^{-1} \nabla_u K(u, c)$. Notice that for flows depending smoothly on a parameter, the Hamiltonian also depends smoothly on it. Consequently, the vector field (4.5) is of the form $B_0 \nabla K(u, c)$. Transforming back to the original variables we obtain $f(y) = B(y) \nabla H(y)$ with $H(y) = K(\vartheta(y))$ (see Corollary 3.6). \square

VII.4.2 Poisson Integrators

The preceding theorem shows that “being a Poisson map and respecting the Casimirs” is characteristic for the flow of a Poisson system. This motivates the following definition.

Definition 4.6. A numerical method $y_1 = \Phi_h(y_0)$ is a *Poisson integrator* for the structure matrix $B(y)$, if the transformation $y_0 \mapsto y_1$ respects the Casimirs and if it is a Poisson map whenever the method is applied to (4.1).

Observe that for a Poisson integrator one has to specify the class of structure matrices $B(y)$. A method will never be a Poisson integrator for all possible $B(y)$.

Example 4.7. The symplectic Euler method reads

$$u_{n+1} = u_n + hu_{n+1}v_n H_v(u_{n+1}, v_n), \quad v_{n+1} = v_n - hu_{n+1}v_n H_u(u_{n+1}, v_n)$$

for the Lotka–Volterra problem (2.13). It produces an excellent long-time behaviour (Fig. 4.1, left picture). We shall show that this is a Poisson integrator for all separable Hamiltonians $H(u, v) = K(u) + L(v)$. For this we compute the Jacobian of the map $(u_n, v_n) \mapsto (u_{n+1}, v_{n+1})$,

$$\begin{pmatrix} 1 - hv_n H_v & 0 \\ hv_n(H_u + u_{n+1}H_{uu}) & 1 \end{pmatrix} \begin{pmatrix} \partial(u_{n+1}, v_{n+1}) \\ \partial(u_n, v_n) \end{pmatrix} = \begin{pmatrix} 1 & hu_{n+1}(H_v + v_n H_{vv}) \\ 0 & 1 - hu_{n+1}H_u \end{pmatrix}$$

(the argument of the partial derivatives of H is (u_{n+1}, v_n) everywhere), and we check in a straightforward fashion the validity of (4.2). A different proof, using differential forms, is given in Sanz-Serna (1994) for a special choice of $H(u, v)$. Similarly, the adjoint of the symplectic Euler method is a Poisson integrator, and so is their composition – the Störmer–Verlet scheme. Composition methods based on this scheme yield high order Poisson integrators, because the composition of Poisson maps is again a Poisson map.

The implicit midpoint rule, though symplectic when applied to canonical Hamiltonian systems, turns out not to be a Poisson map for the structure matrix $B(u, v)$ of (2.13). Figure 4.1 (right picture) shows that the numerical solution does not remain near a closed curve.

It is a difficult task to construct Poisson integrators for general Poisson systems; cf. the overview by Karasözen (2004). First of all, for non-constant $B(y)$ condition (4.2) is no longer a quadratic first integral of the problem augmented by its variational equation (see Sect. VI.4.1). Secondly, the Casimir functions can be arbitrary and we know that only linear and quadratic first integrals can be conserved automatically (Chap. IV). Therefore, Poisson integrators will have to exploit special structures of the particular problem.

Splitting Methods. Consider a (general) Poisson system $\dot{y} = B(y)\nabla H(y)$ and suppose that the Hamiltonian permits a decomposition as $H(y) = H^{[1]}(y) + \dots +$

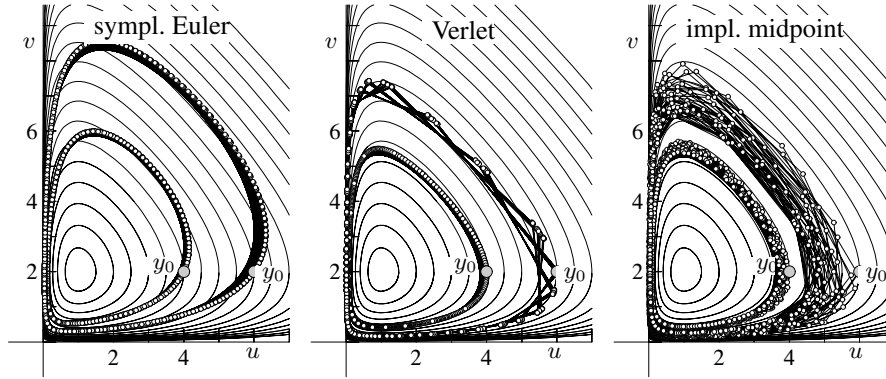


Fig. 4.1. Numerical solutions of the Lotka–Volterra equations (2.13) (step size $h = 0.25$, which is very large compared to the period of the solution; 1000 steps; initial values $(4, 2)$ and $(6, 2)$ for all methods)

$H^{[m]}(y)$, such that the individual systems $\dot{y} = B(y)\nabla H^{[i]}(y)$ can be solved exactly. The flow of these subsystems is a Poisson map and automatically respects the Casimirs, and so does their composition. McLachlan (1993), Reich (1993), and McLachlan & Quispel (2002) present several interesting examples.

Example 4.8. In the previous example of a Lotka–Volterra equation with separable Hamiltonian $H(u, v) = K(u) + L(v)$, the systems with Hamiltonian $K(u)$ and $L(v)$ can be solved explicitly. Since the flow of each of the subsystems is a Poisson map, so is their composition. Combining a half-step with L , a full step with K , and again a half-step with L , we thus obtain the following Verlet-like second-order Poisson integrator:

$$\begin{aligned} u_{n+1/2} &= \exp\left(\frac{h}{2} v_n \nabla L(v_n)\right) u_n \\ v_{n+1} &= \exp\left(-h u_{n+1/2} \nabla K(u_{n+1/2})\right) v_n \\ u_{n+1} &= \exp\left(\frac{h}{2} v_{n+1} \nabla L(v_{n+1})\right) u_{n+1/2}. \end{aligned} \quad (4.6)$$

In the setting of Hamiltonian systems on a manifold, the splitting approach can be formulated in the following way.

Variational Splitting. Consider a Hamiltonian system (2.17) on a symplectic manifold \mathcal{M} , and use a splitting $H = H^{[1]} + H^{[2]}$ of the Hamiltonian in the following algorithm:

1. Let $x_n^+ \in \mathcal{M}$ be the solution at time $h/2$ of the equation for x ,

$$(J\dot{x} - \nabla H^{[1]}(x), \xi) = 0 \quad \text{for all } \xi \in T_x \mathcal{M} \quad (4.7)$$

with initial value $x(0) = x_n$.

2. Let x_{n+1}^- be the solution at time h of

$$(J\dot{x} - \nabla H^{[2]}(x), \xi) = 0 \quad \text{for all } \xi \in T_x \mathcal{M} \quad (4.8)$$

with initial value $x(0) = x_n^+$.

3. Take x_{n+1} as the solution at time $h/2$ of (4.7) with initial value $x(0) = x_{n+1}^-$.

Splitting algorithms for Hamiltonian systems on manifolds have been studied by Dullweber, Leimkuhler & McLachlan (1997) and Benettin, Cherubini & Fassò (2001) in the context of rigid body dynamics; see Sect. VII.5. Lubich (2004) and Faou & Lubich (2004) have studied the above splitting method for applications in quantum molecular dynamics; see Sect. VII.6 for an example.

By Theorem 2.8, the substeps 1.–3. written in coordinates $x = \chi(y)$ are Poisson systems $\dot{y} = B(y)\nabla K^{[i]}(y)$ with $K^{[i]}(y) = H^{[i]}(\chi(y))$, but the algorithm itself is independent of the choice of coordinates. Since the substeps are exact flows of Hamiltonian systems on the manifold \mathcal{M} , their composition yields a symplectic map. In the coordinates y the substeps are the exact flows of Poisson systems, and hence their composition yields a Poisson map.

Poisson Integrators and Symplectic Integrators. Generally we note the following correspondence, which rephrases the remark on symplectic maps and Poisson maps after Definition 4.2. It applies in particular to the symplectic integrators for constrained mechanics of Sect. VII.1.

Lemma 4.9. *An integrator $x_1 = \Psi_h(x_0)$ for a Hamiltonian system (2.17) on a manifold \mathcal{M} is symplectic if and only if the integrator written in local coordinates, $y_1 = \Phi_h(y_0)$ corresponding to a coordinate map $x = \chi(y)$, is a Poisson integrator for the structure matrix $B(y)$ of (2.21).*

VII.4.3 Integrators Based on the Darboux–Lie Theorem

If we explicitly know a transformation $z = \vartheta(y)$ that brings the system $\dot{y} = B(y)\nabla H(y)$ to canonical form (as in Corollary 3.6), we can proceed as follows: compute $z_n = \vartheta(y_n)$; apply a symplectic integrator to the transformed system $\dot{z} = B_0\nabla K(z)$ (B_0 is the matrix (3.18) and $K(z) = H(y)$) which yields $z_{n+1} = \Psi_h(z_n)$; compute finally y_{n+1} from $z_{n+1} = \vartheta(y_{n+1})$. This yields a Poisson integrator by the following lemma.

Lemma 4.10. *Let $z = (u, c) = \vartheta(y)$ be the transformation of Theorem 3.4. Suppose that the integrator $\Phi_h(y)$ takes the form*

$$\Psi_h(z) = \begin{pmatrix} \Psi_h^1(u, c) \\ c \end{pmatrix}$$

in the new variables $z = (u, c)$. Then, $\Phi_h(y)$ is a Poisson integrator if and only if $u \mapsto \Psi_h^1(u, c)$ is a symplectic integrator for every c .

Proof. The integrator $\Phi_h(y)$ is Poisson for the structure matrix $B(y)$ if and only if $\Psi_h(z)$ is Poisson for the matrix B_0 of (3.18); see Exercise 7. By assumption, $\Psi_h(z)$ preserves the Casimirs of B_0 . The identity

$$\Psi'_h(z)B_0\Psi'_h(z)^T = \begin{pmatrix} A J^{-1} A^T & 0 \\ 0 & 0 \end{pmatrix}$$

with $A = \partial\Psi_h^1/\partial u$ proves the statement. \square

Notice that the transformation ϑ has to be global in the sense that it has to be the same for all integration steps. Otherwise a degradation in performance, similar to that of the experiment in Example V.4.3, has to be expected.

Example 4.11. As a first illustration consider the Lotka–Volterra system (2.13). Applying the transformation $\vartheta(u, v) = (\ln u, \ln v) = (p, q)$, this system becomes canonically Hamiltonian with

$$K(p, q) = -H(u, v) = -H(e^p, e^q).$$

If we apply the symplectic Euler method to this Hamiltonian system, and if we transform back to the original variables, we obtain the method

$$\begin{aligned} u_{n+1} &= u_n \exp(h v_n H_v(u_{n+1}, v_n)), \\ v_{n+1} &= v_n \exp(-h u_{n+1} H_u(u_{n+1}, v_n)). \end{aligned} \quad (4.9)$$

In contrast to the method of Example 4.7, (4.9) is also a Poisson integrator for (2.13) if $H(u, v)$ is not separable. If we compose a step with step size $h/2$ of the symplectic Euler method with its adjoint method, then we obtain again, in the case of a separable Hamiltonian, the method (4.6).

Example 4.12 (Ablowitz–Ladik Discrete Nonlinear Schrödinger Equation).

An interesting space discretization of the nonlinear Schrödinger equation is the Ablowitz–Ladik model

$$i \dot{y}_k + \frac{1}{\Delta x^2} (y_{k+1} - 2y_k + y_{k-1}) + |y_k|^2 (y_{k+1} + y_{k-1}) = 0,$$

which we consider under periodic boundary conditions $y_{k+N} = y_k$ ($\Delta x = 1/N$). It is completely integrable (Ablowitz–Ladik 1976) and, as we shall see below, it is a Poisson system with noncanonical Poisson bracket. Splitting the variables into real and imaginary parts, $y_k = u_k + i v_k$, we obtain

$$\begin{aligned} \dot{u}_k &= -\frac{1}{\Delta x^2} (v_{k+1} - 2v_k + v_{k-1}) - (u_k^2 + v_k^2) (v_{k+1} + v_{k-1}) \\ \dot{v}_k &= \frac{1}{\Delta x^2} (u_{k+1} - 2u_k + u_{k-1}) + (u_k^2 + v_k^2) (u_{k+1} + u_{k-1}). \end{aligned}$$

With $u = (u_1, \dots, u_N)$, $v = (v_1, \dots, v_N)$ this system can be written as

$$\begin{pmatrix} \dot{u} \\ \dot{v} \end{pmatrix} = \begin{pmatrix} 0 & -D(u, v) \\ D(u, v) & 0 \end{pmatrix} \begin{pmatrix} \nabla_u H(u, v) \\ \nabla_v H(u, v) \end{pmatrix}, \quad (4.10)$$

where $D = \text{diag}(d_1, \dots, d_N)$ is the diagonal matrix with entries

$$d_k(u, v) = 1 + \Delta x^2(u_k^2 + v_k^2),$$

and the Hamiltonian is

$$H(u, v) = \frac{1}{\Delta x^2} \sum_{l=1}^N (u_l u_{l-1} + v_l v_{l-1}) - \frac{1}{\Delta x^4} \sum_{l=1}^N \ln(1 + \Delta x^2(u_l^2 + v_l^2)).$$

We thus get a Poisson system (the conditions of Lemma 2.3 are directly verified). There are many possibilities to transform this system to canonical form. Tang, Pérez-García & Vázquez (1997) propose the transformation

$$p_k = \frac{1}{\Delta x \sqrt{1 + \Delta x^2 v_k^2}} \arctan\left(\frac{\Delta x}{\sqrt{1 + \Delta x^2 v_k^2}} \cdot u_k\right), \quad q_k = v_k,$$

for which the inverse can be computed in a straightforward way. Here, we suggest the transformation

$$\begin{aligned} p_k &= u_k \sigma(\Delta x^2(u_k^2 + v_k^2)) \\ q_k &= v_k \sigma(\Delta x^2(u_k^2 + v_k^2)) \end{aligned} \quad \text{with} \quad \sigma(x) = \sqrt{\frac{\ln(1+x)}{x}}, \quad (4.11)$$

which treats the variables more symmetrically. Its inverse is

$$\begin{aligned} u_k &= p_k \tau(\Delta x^2(p_k^2 + q_k^2)) \\ v_k &= q_k \tau(\Delta x^2(p_k^2 + q_k^2)) \end{aligned} \quad \text{with} \quad \tau(x) = \frac{\exp x - 1}{x}.$$

Both transformations take the system (4.10) to canonical form. For the transformation (4.11) the Hamiltonian in the new variables is

$$\begin{aligned} H(p, q) &= \frac{1}{\Delta x^2} \sum_{l=1}^N \tau(\Delta x^2(p_l^2 + q_l^2)) \tau(\Delta x^2(p_{l-1}^2 + q_{l-1}^2)) (p_l p_{l-1} + q_l q_{l-1}) \\ &\quad - \frac{1}{\Delta x^2} \sum_{l=1}^N (p_l^2 + q_l^2). \end{aligned}$$

Applying standard symplectic schemes to this Hamiltonian yields Poisson integrators for (4.10).

VII.5 Rigid Body Dynamics and Lie–Poisson Systems

... these topics, which, after all, have occupied workers in geometric mechanics for many years. (R. McLachlan 2003)

An important Poisson system is given by Euler's famous equations for the motion of a rigid body (see left picture of Fig. 5.1), for which we recall the history and derivation and present various structure-preserving integrators. Euler's equations are a particular case of Lie–Poisson systems, which result from a reduction process of Hamiltonian systems on a Lie group.

VII.5.1 History of the Euler Equations

“Le sujet que je me propose de traiter ici, est de la dernière importance dans la Mécanique ; & j’ai déjà fait plusieurs efforts pour le mettre dans tout son jour. Mais, quoique le calcul ait assés bien réussi, & que j’ai découvert des formules analytiques ..., leur application étoit pourtant assujettie à des difficultés qui m’ont paru presque tout à fait insurmontables. Or, depuis que j’ai développé les principes de la connoissance mécanique des corps, la belle propriété des trois axes principaux dont chaque corps est doué, m’a enfin mis en état de vaincre toutes ces difficultés, ...”

(Euler 1758b, p. 154)

A great challenge for Euler were his efforts to establish a mathematical analysis for the motion of a rigid body. Due to the fact that such a body can have an arbitrary shape and mass distribution (see left picture of Fig. 5.2), and that the rotation axis can arbitrarily move with time, the problem is difficult and Euler struggled for many years (all these articles are collected in *Opera Omnia*, Ser. II, Vols. 7 and 8). The breakthrough was enabled by the discovery that any body, as complicated as may be its configuration, reduces to an inertia ellipsoid with three principal axes and three numbers, the principal moments of inertia (Euler 1758a; see the middle picture of Fig. 5.2 and the citation).

$$\begin{aligned} dx + \frac{cc - bb}{aa} \cdot yz dt &= \frac{2gPdt}{Ma a} \\ dy + \frac{aa - cc}{bb} \cdot xz dt &= \frac{2gQdt}{Mb b} \\ dz + \frac{bb - aa}{cc} \cdot xy dt &= \frac{2gRdt}{Mc c} \end{aligned}$$

$$\alpha p' = A p' - H q' - G r';$$

, la quatrième et la si
' , on aura pareillement

$$\alpha q' = B q' - F r' - H p';$$

me, la cinquième et la
' , q' , on aura

$$\alpha r' = C r' - G p' - F q';$$

Fig. 5.1. Left picture: first publication of the Euler equations in Euler (1758b). Right picture: principal axes as eigenvectors in Lagrange (1788)

The Inertia Ellipsoid. We choose a moving coordinate system connected to the body \mathcal{B} and we consider motions of the body where the origin is fixed. By another of Euler’s famous theorems, any such motion is infinitesimally a rotation around an axis. We represent the rotation axis of the body by the *direction* of a vector ω and the speed of rotation by the *length* of ω . Then the velocity of a mass point x of \mathcal{B} is given by the exterior product

$$v = \omega \times x = \begin{pmatrix} \omega_2 x_3 - \omega_3 x_2 \\ \omega_3 x_1 - \omega_1 x_3 \\ \omega_1 x_2 - \omega_2 x_1 \end{pmatrix} = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad (5.1)$$

(orthogonal to ω , orthogonal to x , and of length $\|\omega\| \cdot \|x\| \cdot \sin \gamma$; see the left picture of Fig. 5.2). The *kinetic energy* is obtained by integrating the energy of the mass

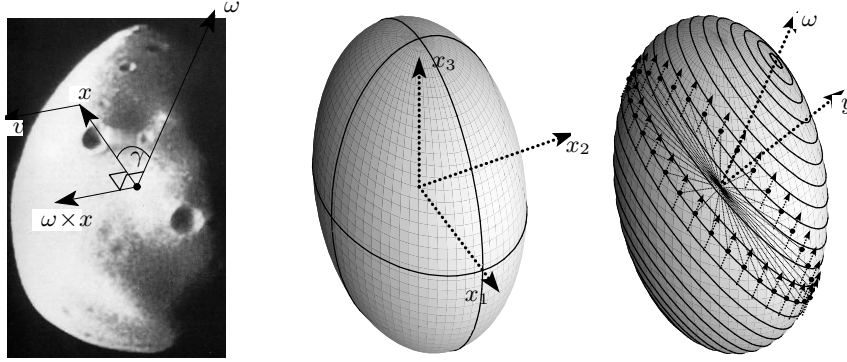


Fig. 5.2. A rigid body rotating around a variable axis (left); the corresponding inertia ellipsoid (middle); the corresponding angular momentum (right)

points dm over the body

$$\begin{aligned} T &= \frac{1}{2} \int_B \|\omega \times x\|^2 dm \\ &= \frac{1}{2} \int_B \left((\omega_2 x_3 - \omega_3 x_2)^2 + (\omega_3 x_1 - \omega_1 x_3)^2 + (\omega_1 x_2 - \omega_2 x_1)^2 \right) dm. \end{aligned} \quad (5.2)$$

If this is multiplied out, one obtains

$$T = \frac{1}{2} \omega^T \Theta \omega, \text{ where } \Theta_{ii} = \int_B (x_k^2 + x_\ell^2) dm, \Theta_{ik} = - \int_B x_i x_k dm, \quad (i \neq k, \ell). \quad (5.3)$$

Euler (1758a) showed, by endless trigonometric transformations, that there exist principal axes of the body in which this expression takes the form

$$T = \frac{1}{2} \left(I_1 \omega_1^2 + I_2 \omega_2^2 + I_3 \omega_3^2 \right). \quad (5.4)$$

This was historically the first transformation of such a 3×3 quadratic form to diagonal form. Later, Lagrange (1788) discovered that these axes were the eigenvectors of the matrix Θ and the moments of inertia I_k the corresponding eigenvalues (without calling them so, see the right picture of Fig. 5.1).

The Angular Momentum. The first law of Newton's *Principia* states that the *momentum* $v \cdot m$ of a mass point remains constant in the absence of exterior forces. The corresponding quantity for *rotational* motion is the *angular momentum*, i.e., the exterior product $x \times v$ times the mass. Integrating over the body we obtain, with (5.1),

$$y = \int_B (x \times v) dm = \int_B \left(x \times (\omega \times x) \right) dm. \quad (5.5)$$

If this is multiplied out, the matrix Θ appears again and one obtains the surprising result (due to Poinsot 1834)

$$y = \Theta \omega, \quad \text{or, in the principal axes coordinates,} \quad y_k = I_k \omega_k. \quad (5.6)$$

Such a relation is familiar from the theory of conjugate diameters (Apollonius, Book II, Prop. VI): the angular momentum is a vector orthogonal to the plane of vectors conjugate to ω (see the right picture of Fig. 5.2).

The Euler Equations. Euler’s paper (1758a), on his discovery of the principal axes, is immediately followed by Euler (1758b), where he derives his equations for the motion of a rigid body by long, doubtful and often criticized calculations, repeated in a little less doubtful manner in Euler’s monumental treatise (1765). Beauty and elegance, not only of the result, but also of the proof, is due to Poincot (1834) and Hayward (1856). It is masterly described by Klein & Sommerfeld (1897), and in Chapter 6 of Arnold (1989).

From now on we choose the coordinate system, moving with the body, such that the inertia tensor remains diagonal. We also watch the motion of the body from a coordinate system stationary in the space. The transformation of a vector $x \in \mathbb{R}^3$ in the body frame ⁴, to the corresponding $\tilde{x} \in \mathbb{R}^3$ in the stationary frame, is denoted by

$$\tilde{x} = Q(t)x. \quad (5.7)$$

The matrix $Q(t)$ is orthogonal and describes the motion of the body: for $x = e_i$ we see that the columns of $Q(t)$ are the coordinates of the body’s principal axes in the stationary frame.

The analogous statement to Newton’s first law for rotational motion is: *in the absence of exterior angular forces, the angular momentum \tilde{y} , seen from the fixed coordinate system, is a constant vector* ⁵. This same vector y , seen from the moving frame, which at any instance rotates with the body around the vector ω , rotates in the *opposite* direction. Therefore we have from (5.1), by changing the signs of ω , the derivatives

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & \omega_3 & -\omega_2 \\ -\omega_3 & 0 & \omega_1 \\ \omega_2 & -\omega_1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}. \quad (5.8)$$

If we insert $\omega_k = y_k/I_k$ from (5.6), we obtain

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & y_3/I_3 & -y_2/I_2 \\ -y_3/I_3 & 0 & y_1/I_1 \\ y_2/I_2 & -y_1/I_1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} (I_3^{-1} - I_2^{-1}) y_3 y_2 \\ (I_1^{-1} - I_3^{-1}) y_1 y_3 \\ (I_2^{-1} - I_1^{-1}) y_2 y_1 \end{pmatrix} \quad (5.9)$$

or, by rearranging the products the other way round,

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix} \begin{pmatrix} y_1/I_1 \\ y_2/I_2 \\ y_3/I_3 \end{pmatrix}, \quad (5.10)$$

⁴ Long-standing tradition, from Klein to Arnold, uses capitals for denoting the coordinates in this moving frame; but this would lead to confusion with our subsequent matrix notation

⁵ For a proof of this statement by d’Alembert’s Principle, see Sommerfeld (1942), §II.13.

written in two different ways as a Poisson system, whose right hand vectors are the gradients of $C(y) = \frac{1}{2} \sum_{k=1}^3 y_k^2$ and $H(y) = \frac{1}{2} \sum_{k=1}^3 I_k^{-1} y_k^2$, respectively. These are the two quadratic invariants of Chap. IV. The first represents the length of the constant angular momentum \tilde{y} in the orthogonal body frame, and the second represents the energy (5.4).

Computation of the Position Matrix $Q(t)$. Once we have solved the Euler equations for $y(t)$, we obtain the rotation vector $\omega(t)$ by (5.6). It remains to find the matrix $Q(t)$ which gives the position of our rotating body. We know that the columns of the matrix Q , seen in the stationary frame, correspond to the unit vectors e_i in the body frame. These rotate, by (5.1), with the velocity

$$(\omega \times e_1, \omega \times e_2, \omega \times e_3) = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix} =: W. \quad (5.11)$$

We thus obtain \dot{Q} , the rotational velocity expressed in the stationary frame, by the back transformation (5.7):

$$\dot{Q} = QW \quad \text{or} \quad Q^T \dot{Q} = W. \quad (5.12)$$

This is a differential system for Q which, because W is skew-symmetric, preserves the orthogonality of Q . The problem is solved – in theory.

VII.5.2 Hamiltonian Formulation of Rigid Body Motion

In order to open the door for efficient numerical algorithms, we treat the rigid body as a constrained Hamiltonian system.

Position Variables. The position of the rigid body at time t is determined, in view of (5.7), by a three-dimensional orthogonal matrix $Q(t)$. The constraints to be respected are thus $Q^T Q - I = 0$.

Kinetic Energy. As in (5.12), we associate with Q and \dot{Q} the skew-symmetric matrix $W = Q^T \dot{Q}$ whose entries ω_k , arranged as in (5.11), determine the kinetic energy by (5.4):

$$T = \frac{1}{2} (I_1 \omega_1^2 + I_2 \omega_2^2 + I_3 \omega_3^2).$$

For any diagonal matrix $D = \text{diag}(d_1, d_2, d_3)$ we observe

$$\text{trace}(WDW^T) = (d_2 + d_3)\omega_1^2 + (d_3 + d_1)\omega_2^2 + (d_1 + d_2)\omega_3^2.$$

Therefore, with

$$I_1 = d_2 + d_3, \quad I_2 = d_3 + d_1, \quad I_3 = d_1 + d_2 \quad \text{or} \quad d_k = \int_{\mathcal{B}} x_k^2 dm \quad (5.13)$$

(note that $d_k > 0$ for all bodies that have interior points), we obtain the kinetic energy as

$$T = \frac{1}{2} \text{trace} (W D W^T). \quad (5.14)$$

Inserting $W = Q^T \dot{Q}$, we have

$$T = \frac{1}{2} \text{trace} (Q^T \dot{Q} D \dot{Q}^T Q) = \frac{1}{2} \text{trace} (\dot{Q} D \dot{Q}^T), \quad (5.15)$$

since Q is an orthogonal matrix.

Conjugate Variables. We now have an expression for the kinetic energy in terms of derivatives of position coordinates and are able to introduce the conjugate momenta

$$P = \partial T / \partial \dot{Q} = \dot{Q} D. \quad (5.16)$$

If we suppose to have, in addition to T , a potential $U(Q)$, we get the Hamiltonian

$$H(P, Q) = \frac{1}{2} \text{trace} (P D^{-1} P^T) + U(Q). \quad (5.17)$$

Lagrange Multipliers. The constraints are given by the orthogonality of Q , i.e., the equation $g(Q) = Q^T Q - I = 0$. Since this matrix is always symmetric, this consists of $\frac{1}{2}n(n+1) = 6$ independent algebraic conditions, calling for six Lagrange multipliers. If the expression $G(Q)^T \lambda$ in (1.9) is actually computed, it turns out that this term becomes the product $Q \Lambda$, where the six Lagrange multipliers are arranged in a symmetric matrix Λ ; see also formula (IV.9.6). Thus, the constrained Hamiltonian system (1.9) reads in our case, with $\nabla U = (\partial U / \partial Q_{ij})$,

$$\begin{aligned} \dot{Q} &= P D^{-1} \\ \dot{P} &= -\nabla U(Q) - Q \Lambda \quad (\Lambda \text{ symmetric}) \\ 0 &= Q^T Q - I. \end{aligned} \quad (5.18)$$

Reduction to the Euler Equations. The key idea is to introduce the matrix

$$Y = Q^T P = Q^T \dot{Q} D = W D = \begin{pmatrix} 0 & -d_2 \omega_3 & d_3 \omega_2 \\ d_1 \omega_3 & 0 & -d_3 \omega_1 \\ -d_1 \omega_2 & d_2 \omega_1 & 0 \end{pmatrix}, \quad (5.19)$$

where the ω_k can be further expressed in terms of the angular momenta $y_k = I_k \omega_k$. Using the notation $\text{skew}(A) = \frac{1}{2}(A - A^T)$, we see, with (5.13), that

$$Y - Y^T = 2 \text{skew}(Y) = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix} \quad (5.20)$$

contains just the angular momenta. Moreover, DY is skew-symmetric. By (5.18) the derivative of Y is seen to be

$$\dot{Y} = \dot{Q}^T P + Q^T \dot{P} = D^{-1} P^T P - Q^T \nabla U(Q) - \Lambda = D^{-1} Y^T Y - Q^T \nabla U(Q) - \Lambda.$$

Taking the skew-symmetric part of this equation, the symmetric matrix Λ drops out and we obtain

$$\text{skew}(\dot{Y}) = \text{skew}(D^{-1} Y^T Y) - \text{skew}(Q^T \nabla U(Q)). \quad (5.21)$$

These are, for $U = 0$, precisely the above Euler equations, obtained a second time.

VII.5.3 Rigid Body Integrators

For a numerical simulation of rigid body motions, one can either solve the constrained Hamiltonian system (5.18), or one can solve the differential equation (5.21) for the angular momentum $Y(t)$ in tandem with the equation (5.12) for $Q(t)$. We consider the following approaches: (I) an efficient application of the RATTLE algorithm (1.26), and (II) various splitting methods.

(I) RATTLE

We apply the symplectic RATTLE algorithm (1.26) to the system (5.18), and rewrite the formulas in terms of the variables Y and Q . This approach has been proposed and developed independently by McLachlan & Scovel (1995) and Reich (1994).

An application of the RATTLE algorithm (1.26) to the system (5.18) yields

$$\begin{aligned} P_{1/2} &= P_0 - \frac{h}{2} \nabla U(Q_0) - \frac{h}{2} Q_0 \Lambda_0 \\ Q_1 &= Q_0 + h P_{1/2} D^{-1}, \quad Q_1^T Q_1 = I \\ P_1 &= P_{1/2} - \frac{h}{2} \nabla U(Q_1) - \frac{h}{2} Q_1 \Lambda_1, \quad Q_1^T P_1 D^{-1} + D^{-1} P_1^T Q_1 = 0, \end{aligned} \quad (5.22)$$

where both Λ_0 and Λ_1 are symmetric matrices. We let $Y_0 = Q_0^T P_0$, $Y_1 = Q_1^T P_1$, and $Z = Q_0^T P_{1/2} D^{-1}$. We multiply the first relation of (5.22) by Q_0^T , the last relation by Q_1^T , and we eliminate the symmetric matrices Λ_0 and Λ_1 by taking the skew-symmetric parts of the resulting equations. The orthogonality of $Q_0^T Q_1 = I + hZ$ implies $hZ^T Z = -(Z + Z^T)$, which can then be used to simplify the last relation. Altogether this results in the following algorithm.

Algorithm 5.1. Let Q_0 be orthogonal and DY_0 be skew-symmetric. One step $(Q_0, Y_0) \mapsto (Q_1, Y_1)$ of the method then reads as follows:

– find Z such that $I + hZ$ is orthogonal and

$$\text{skew}(ZD) = \text{skew}(Y_0) - \frac{h}{2} \text{skew}(Q_0^T \nabla U(Q_0)), \quad (5.23)$$

– compute $Q_1 = Q_0(I + hZ)$,

– compute Y_1 such that DY_1 is skew-symmetric and

$$\text{skew}(Y_1) = \text{skew}(ZD) - \text{skew}((Z + Z^T)D) - \frac{h}{2} \text{skew}(Q_1^T \nabla U(Q_1)).$$

The second step is explicit, and the third step represents a linear equation for the elements of Y_1 .

Computation of the First Step. We write for the known part of equation (5.23)

$$\text{skew}(Y_0) - \frac{h}{2} \text{skew}(Q_0^T \nabla U(Q_0)) = \begin{pmatrix} 0 & -\alpha_3 & \alpha_2 \\ \alpha_3 & 0 & -\alpha_1 \\ -\alpha_2 & \alpha_1 & 0 \end{pmatrix} = A \quad (5.24)$$

and have to solve

$$\frac{1}{2}(ZD - DZ^T) = A, \quad (I + hZ^T)(I + hZ) = I, \quad \frac{1}{2}(ZD + DZ^T) = S$$

(the trick was to add the last equation with S an unknown symmetric matrix). Elimination gives $Z = (A + S)D^{-1}$ and $Z^T = D^{-1}(S - A)$. Both inserted into the second equation lead to a Riccati equation for S . There exist efficient algorithms for such problems; see the reference in Sect. IV.5.3 and a detailed explanation in McLachlan & Zanna (2005).

Remark 5.2 (Moser–Veselov Algorithm). An independent access to the above formulas is given in a remarkable paper by Moser & Veselov (1991), by treating the rigid body through a *discretized* variational principle, similar to the ideas of Sect. VI.6.2. The equivalence is explained by McLachlan & Zanna (2005), following a suggestion of B. Leimkuhler and S. Reich.

Quaternions (Euler Parameters). An efficient implementation of the above algorithm requires suitable representations of orthogonal matrices, and the use of quaternions is a standard approach.

After having revolutionized Lagrangian mechanics (see Chapt. VI), Hamilton struggled for years to generalize complex analysis to three dimensions. He finally achieved his dream, however not in three dimensions, but in *four*, and founded in 1843 the theory of quaternions.

For an introduction to quaternions (whose coefficients are sometimes called Euler parameters) we refer to Sects. IV.2 and IV.3 of Klein (1908), and for their use in numerical simulations to Sects. 9.3 and 11.3 of Haug (1989). Quaternions can be written as $e = e_0 + ie_1 + je_2 + ke_3$, where multiplication is defined via the relations $i^2 = j^2 = k^2 = -1$, $ij = k$, $jk = i$, $ki = j$, and $ji = -k$, $kj = -i$, $ik = -j$. The product of two quaternions $e \cdot f$ (written in matrix notation) is

$$(e_0 + ie_1 + je_2 + ke_3) \cdot (f_0 + if_1 + jf_2 + kf_3) = \begin{pmatrix} e_0 & -e_1 & -e_2 & -e_3 \\ e_1 & e_0 & -e_3 & e_2 \\ e_2 & e_3 & e_0 & -e_1 \\ e_3 & -e_2 & e_1 & e_0 \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \end{pmatrix} \quad (5.25)$$

We see (in grey) that in dimensions 1, 2, 3 appears a skew-symmetric matrix E whose structure is familiar to us. This part of the matrix changes sign if the two factors are permuted.

An important discovery, for three dimensional applications of the quaternions, is the following: if a quaternion p is a 3-vector (i.e., has $p_0 = 0$), then $p' = e \cdot p \cdot \bar{e}$ is a 3-vector, too, and the map $p \mapsto p'$ is described by the matrix

$$Q(e) = \|e\|^2 I + 2e_0 E + 2E^2, \quad E = \begin{pmatrix} 0 & -e_3 & e_2 \\ e_3 & 0 & -e_1 \\ -e_2 & e_1 & 0 \end{pmatrix} \quad (5.26)$$

where $\bar{e} = e_0 - ie_1 - je_2 - ke_3$ and $\|e\|^2 = e \cdot \bar{e} = e_0^2 + e_1^2 + e_2^2 + e_3^2$.

Lemma 5.3. *If $\|e\| = 1$, then the matrix $Q(e)$ is orthogonal. Every orthogonal matrix with $\det Q = 1$ can be written in this form. We have $Q(e)Q(f) = Q(ef)$, so that the multiplication of orthogonal matrices corresponds to the multiplication of quaternions.*

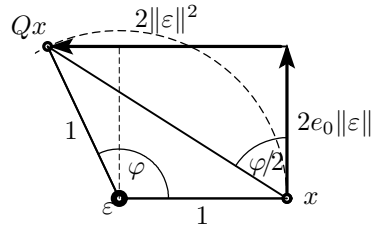
Geometrically, the matrix Q effects a rotation around the axis $\varepsilon = (e_1, e_2, e_3)^T$ with rotation angle φ which satisfies $\tan(\varphi/2) = \|\varepsilon\|/e_0$.

Proof. The condition $Q^T Q = I$ can be verified directly using $E^T = -E$ and $E^3 = -(e_1^2 + e_2^2 + e_3^2)E$. The reciprocal statement is a famous theorem of Euler; it is based on the fact that ε is an eigenvector of Q , which in dimension 3×3 always exists. The formula for $Q(e)Q(f)$ follows from $e \cdot f \cdot p \cdot \bar{f} \cdot \bar{e} = (e \cdot f) \cdot p \cdot (\bar{e} \cdot \bar{f})$.

The geometric property follows from the virtues of the exterior product, because by (5.1) the matrix Q maps a vector x to

$$x + 2e_0 \varepsilon \times x + 2 \varepsilon \times (\varepsilon \times x).$$

This consists in a rectangular movement in a plane orthogonal to ε ; first vertical to x by an amount $2e_0 \|\varepsilon\|$ (times the distance of x), then parallel to x by an amount $2\|\varepsilon\|^2$.



Applying Pythagoras' Theorem as $(2e_0 \|\varepsilon\|)^2 + (2\|\varepsilon\|^2 - 1)^2 = 1$, it turns out that the map is norm preserving if $e_0^2 + \|\varepsilon\|^2 = 1$. The angle $\varphi/2$, whose tangens can be seen to be $\|\varepsilon\|/e_0$, is an angle at the circumference of the circle for the rotation angle φ at the center. \square

For an efficient implementation of Algorithm 5.1 we represent the orthogonal matrices Q_0 , Q_1 , and $I + hZ$ by quaternions. This reduces the dimension of the systems, and step 2 becomes a simple multiplication of quaternions. For solving the nonlinear system of step 1, we let $I + hZ = Q(e)$. With the values of α_i from (5.24) and with $\text{skew}(hZD) = 2e_0 \text{skew}(ED) + 2 \text{skew}(E^2D)$, the equation (5.23) becomes

$$\begin{pmatrix} h\alpha_1 \\ h\alpha_2 \\ h\alpha_3 \end{pmatrix} = 2e_0 \begin{pmatrix} I_1 e_1 \\ I_2 e_2 \\ I_3 e_3 \end{pmatrix} + 2 \begin{pmatrix} (I_3 - I_2)e_2 e_3 \\ (I_1 - I_3)e_3 e_1 \\ (I_2 - I_1)e_1 e_2 \end{pmatrix}, \quad (5.27)$$

which, together with $e_0^2 + e_1^2 + e_2^2 + e_3^2 = 1$, represent four quadratic equations for four unknowns. We solve them very quickly by a few fixed-point iterations: update

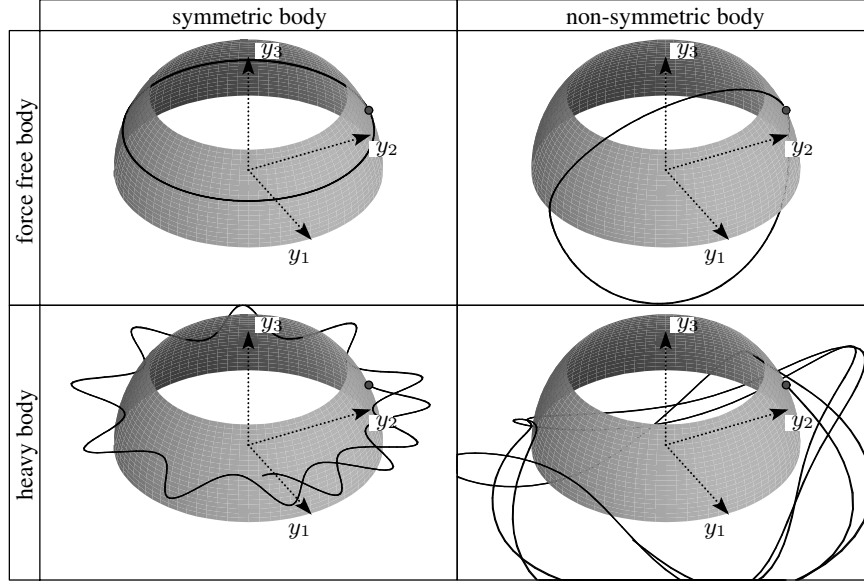


Fig. 5.3. Numerical solutions of the rigid body equations; without/with gravitation, with/without symmetry. Initial values $y_{10} = 0.2, y_{20} = 1.0, y_{30} = 0.4$; initial position of Q_0 determined by the quaternion $e_0 = 0.4, e_1 = 0.2, e_2 = 0.4, e_3 = 0.8$; moments of inertia $I_1 = 0.5, I_2 = 0.85$ (0.5 in the symmetric case), $I_3 = 1$; step size $h = 0.1$, integration interval $0 \leq t \leq 30$

successively e_i from the i th equation of (5.27) and then e_0 from the normalization condition. A Fortran subroutine RATORI for this algorithm is available on the homepage <http://www.unige.ch/~hairer>.

Conservation of Casimir and Hamiltonian. It is interesting to note that, in the absence of a potential, the Algorithm 5.1 preserves exactly the Casimir $y_1^2 + y_2^2 + y_3^2$ and, more surprisingly, also the Hamiltonian $\frac{1}{2}(y_1^2/I_1 + y_2^2/I_2 + y_3^2/I_3)$. This can be seen as follows: without any potential we have $\text{skew}(Y_0) = \text{skew}(ZD)$ and $\text{skew}(Y_1) = -\text{skew}(Z^T D)$, so that the vectors $(y_{10}, y_{20}, y_{30})^T$ and $(y_{11}, y_{21}, y_{31})^T$ are equal to $u + v$ and $u - v$, respectively, where u and v are the vectors of the right-hand side of (5.27). Since u and v are orthogonal, we have $\|u + v\| = \|u - v\|$, which proves the conservation of the Casimir.

To prove the conservation of the Hamiltonian, we first multiply the relation (5.27) with $G = \text{diag}(1/\sqrt{I_1}, 1/\sqrt{I_2}, 1/\sqrt{I_3})$, and then apply the same arguments. The vectors Gu and Gv are still orthogonal.

Example 5.4 (Force-Free and Heavy Top). We present in Fig. 5.3 the numerical solutions y_i obtained by the above algorithm. In the case of the heavy top, we assume the centre of gravity to be $(0, 0, 1)$ in the body frame, and assume that the third coordinate of the stationary frame is vertical. The potential energy due to gravity is

then given by $U(Q) = q_{33}$ and, expressed by quaternions (5.26), it is $U = e_0^2 - e_1^2 - e_2^2 + e_3^2$.

(II) Splitting Methods

As before we consider the differential equation (5.21) for the angular momenta in the body y_1, y_2, y_3 together with the differential equation (5.12) for the rotation matrix Q . An obvious splitting in the presence of a potential is

$$\varphi_{h/2}^U \circ \Phi_h^T \circ \varphi_{h/2}^U, \quad (5.28)$$

where φ_t^U represents the exact flow of

$$\dot{Q} = 0, \quad \text{skew}(\dot{Y}) = -\text{skew}(Q^T \nabla U(Q)),$$

and Φ_h^T is a suitable numerical approximation of the system corresponding to kinetic energy only, i.e., without any potential $U(Q)$. The use of splitting techniques for rigid body dynamics was proposed by Touma & Wisdom (1994), McLachlan (1993), Reich (1994), and Dullweber, Leimkuhler & McLachlan (1997). Fassò (2003) presents a careful study and comparison of various ways of splitting the kinetic energy.

Computation of Φ_h^T . We do this by splitting once again, by letting several moments of inertia tending to infinity (and the corresponding ω_i tending to zero). In order to avoid formal difficulties with infinite denominators, we write the system (5.10) together with (5.12) in the form

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix} \begin{pmatrix} T_{y_1}(y) \\ T_{y_2}(y) \\ T_{y_3}(y) \end{pmatrix} \quad (5.29)$$

$$\dot{Q} = Q \begin{pmatrix} 0 & -T_{y_3}(y) & T_{y_2}(y) \\ T_{y_3}(y) & 0 & -T_{y_1}(y) \\ -T_{y_2}(y) & T_{y_1}(y) & 0 \end{pmatrix}, \quad (5.30)$$

where $T(y) = \frac{1}{2}(y_1^2/I_1 + y_2^2/I_2 + y_3^2/I_3)$ is the kinetic energy, and $T_{y_i}(y)$ denote partial derivatives.

Three Rotations Splitting. An obvious splitting of the kinetic energy is

$$T(y) = R_1(y) + R_2(y) + R_3(y), \quad R_i(y) = y_i^2/(2I_i), \quad (5.31)$$

which results in the numerical method

$$\Phi_h^T = \varphi_{h/2}^{R_3} \circ \varphi_{h/2}^{R_2} \circ \varphi_h^{R_1} \circ \varphi_{h/2}^{R_2} \circ \varphi_{h/2}^{R_3},$$

where $\varphi_t^{R_i}$ is the exact flow of (5.29)-(5.30) with $T(y)$ replaced by $R_i(y)$. The flow $\varphi_t^{R_1}$ is easily obtained: y_1 remains constant and the second and third equation in (5.29) boil down to the harmonic oscillator. We obtain

$$y(t) = S(\alpha t)y(0), \quad Q(t) = Q(0)S(\alpha t)^T \quad (5.32)$$

with $\alpha = y_1(0)/I_1$ and the rotation matrix

$$S(\theta) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & \sin \theta \\ 0 & -\sin \theta & \cos \theta \end{pmatrix}.$$

Similar simple formulas are obtained for the exact flows corresponding to R_2 and R_3 .

Symmetric + Rotation Splitting. It is often advantageous, in particular for a nearly symmetric body ($I_1 \approx I_2$), to consider the splitting

$$T(y) = R(y) + S(y), \quad R(y) = \left(\frac{1}{I_1} - \frac{1}{I_2} \right) \frac{y_1^2}{2}, \quad S(y) = \frac{1}{2} \left(\frac{y_1^2 + y_2^2}{I_2} + \frac{y_3^2}{I_3} \right)$$

and the corresponding numerical integrator

$$\Phi_h^T = \varphi_{h/2}^R \circ \varphi_h^S \circ \varphi_{h/2}^R.$$

The exact flow φ_t^R is the same as (5.32) with I_1^{-1} replaced by $I_1^{-1} - I_2^{-1}$. The flow of the symmetric force-free top φ_t^S possesses simple analytic formulas, too (see the first picture of Fig. 5.3): we observe a precession of y with constant speed around a cone and a rotation of the body around ω with constant speed. Therefore

$$y(t) = B(\beta t)y(0), \quad Q(t) = Q(0)A(t)B(\beta t)^T, \quad (5.33)$$

where $\beta = (I_3^{-1} - I_2^{-1})y_3(0)$, and

$$A(t) = \exp \left(\frac{t}{I_2} \begin{pmatrix} 0 & -y_3(0) & y_2(0) \\ y_3(0) & 0 & -y_1(0) \\ -y_2(0) & y_1(0) & 0 \end{pmatrix} \right), \quad B(\theta) = \begin{pmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

This can also be checked directly by differentiation.

Similar to the previous algorithm it is advantageous to use a representation of the appearing orthogonal matrices by quaternions. The correspondence between the orthogonal rotation matrices appearing in (5.32) and (5.33) and their quaternions is, in accordance with Lemma 5.3, the following:

$$\begin{aligned} S(\theta)^T &\leftrightarrow \cos(\theta/2) + i \sin(\theta/2) \\ B(\theta)^T &\leftrightarrow \cos(\theta/2) + k \sin(\theta/2) \\ A(t) &\leftrightarrow \cos(\vartheta/2) + a^{-1} \sin(\vartheta/2) (i y_1(0) + j y_2(0) + k y_3(0)), \end{aligned}$$

where $a = \sqrt{y_1(0)^2 + y_2(0)^2 + y_3(0)^2}$ and $\vartheta = at/I_2$. The matrix multiplications in the algorithm can therefore be done very efficiently. A Fortran subroutine QUATER for the “symmetric + rotation splitting” algorithm is available on the homepage <<http://www.unige.ch/~hairer>>.

VII.5.4 Lie–Poisson Systems

In Sect. VII.5.1 we have seen that the reduction of the equations of motion of the rigid body leads to the Poisson system (5.10) with a structure matrix whose entries are linear functions. Here we consider more general Poisson systems

$$\dot{y} = B(y)\nabla H(y), \quad (5.34)$$

where the structure matrix $B(y)$ depends linearly on y , i.e.,

$$b_{ij}(y) = \sum_{k=1}^n C_{ji}^k y_k \quad \text{for } i, j = 1, \dots, n. \quad (5.35)$$

Such systems, called Lie–Poisson systems, are closely related to differential equations on the dual of Lie algebras; see Marsden & Ratiu (1999), Chaps. 13 and 14, for an in-depth discussion of this theory.

Recall that a Lie algebra is a vector space with a bracket which is anti-symmetric and satisfies the Jacobi identity (Sect. IV.6). Let E_1, \dots, E_n be a basis of a vector space, and define a bracket by

$$[E_i, E_j] = \sum_{k=1}^n C_{ij}^k E_k \quad (5.36)$$

with C_{ij}^k from (5.35). If the structure matrix $B(y)$ of (5.35) is skew-symmetric and satisfies (2.10), then this bracket makes the vector space a Lie algebra (the verification is left as an exercise). The coefficients C_{ij}^k are called *structure constants* of the Lie algebra. Conversely, if we start from a Lie algebra with bracket given by (5.36), then the matrix $B(y)$ defined by (5.35) is the structure matrix of a Poisson bracket.

Let \mathfrak{g} be a Lie algebra with a basis E_1, \dots, E_n , and let \mathfrak{g}^* be the dual of the Lie algebra, i.e., the vector space of all linear forms $Y : \mathfrak{g} \rightarrow \mathbb{R}$. The duality is written as $\langle Y, X \rangle$ for $Y \in \mathfrak{g}^*$ and $X \in \mathfrak{g}$. We denote by F_1, \dots, F_n the dual basis defined by $\langle F_i, E_j \rangle = \delta_{ij}$, the Kronecker δ .

Theorem 5.5. *Let \mathfrak{g} be a Lie algebra with basis E_1, \dots, E_n satisfying (5.36). To $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ we associate $Y = \sum_{j=1}^n y_j F_j \in \mathfrak{g}^*$, and we consider a Hamiltonian⁶ $H(y) = H(Y)$.*

Then, the Poisson system $\dot{y} = B(y)\nabla H(y)$ with $B(y)$ given by (5.35) is equivalent to the following differential equation on the dual \mathfrak{g}^ :*

$$\langle \dot{Y}, X \rangle = \langle Y, [H'(Y), X] \rangle \quad \text{for all } X \in \mathfrak{g}, \quad (5.37)$$

where $H'(Y) = \sum_{j=1}^n \frac{\partial H(y)}{\partial y_j} E_j$.

⁶ We use the same symbol H for the functions $H : \mathbb{R}^n \rightarrow \mathbb{R}$ and $H : \mathfrak{g}^* \rightarrow \mathbb{R}$.

Proof. Differentiating $H(y) = H(Y)$ with respect to y_i gives

$$\frac{\partial H(y)}{\partial y_i} = \langle F_i, H'(Y) \rangle \quad \text{and} \quad H'(Y) = \sum_{j=1}^n \frac{\partial H(y)}{\partial y_j} E_j.$$

Here we have used the identification $(\mathfrak{g}^*)^* = \mathfrak{g}$, because $H'(Y)$ is actually an element of $(\mathfrak{g}^*)^*$. With this formula for $H'(Y)$ we are able to compute

$$\langle Y, [H'(Y), E_i] \rangle = \left\langle Y, \sum_{j=1}^n \frac{\partial H(y)}{\partial y_j} [E_j, E_i] \right\rangle = \sum_{j=1}^n \sum_{k=1}^n \frac{\partial H(y)}{\partial y_j} C_{ji}^k \langle Y, E_k \rangle,$$

where we have used (5.36). Since $\langle \dot{Y}, E_i \rangle = \dot{y}_i$ and $\langle Y, E_k \rangle = y_k$, this shows that the differential equation (5.37) is equivalent to

$$\dot{y}_i = \sum_{j=1}^n \left(\sum_{k=1}^n C_{ji}^k y_k \right) \frac{\partial H(y)}{\partial y_j},$$

which is nothing more than $\dot{y} = B(y) \nabla H(y)$ with $B(y)$ from (5.35). \square

We remark that (5.37) can be reformulated as

$$\dot{Y} = \text{ad}_{H'(Y)}^* Y,$$

where ad_A^* is the adjoint of the operator $\text{ad}_A(X) = [A, X]$.

Equation (5.37) is similar in appearance to the Lie bracket equation $\dot{L} = [A(L), L] = \text{ad}_{A(L)} L$ of Sect. IV.3.2. When \mathfrak{g} is the Lie algebra of a matrix Lie group G , then solutions of that equation are of the form $L(t) = \text{Ad}_{U(t)} L_0$ where

$$\text{Ad}_U X = UXU^{-1}; \quad (5.38)$$

see the proof of Lemma IV.3.4. Similarly, for the solution of (5.37) we have the following.

Theorem 5.6. *Consider a matrix Lie group G with Lie algebra \mathfrak{g} . Then, the solution $Y(t) \in \mathfrak{g}^*$ of (5.37) with initial value $Y_0 \in \mathfrak{g}^*$ is given by*

$$\langle Y(t), X \rangle = \langle Y_0, U(t)^{-1} X U(t) \rangle \quad \text{for all } X \in \mathfrak{g}, \quad (5.39)$$

where $U(t) \in G$ satisfies

$$\dot{U} = -H'(Y(t))U, \quad U(0) = I. \quad (5.40)$$

Equation (5.39) can be written as

$$Y(t) = \text{Ad}_{U(t)^{-1}}^* Y_0,$$

where $\text{Ad}_{U^{-1}}^*$ is the adjoint of $\text{Ad}_{U^{-1}}$. The solution $Y(t)$ of (5.37) thus lies on the *coadjoint orbit*

$$Y(t) \in \{\text{Ad}_{U^{-1}}^* Y_0; U \in G\}. \quad (5.41)$$

In coordinates $Y = \sum_{j=1}^n y_j F_j$, we note $y_j = \langle Y_0, U(t)^{-1} E_j U(t) \rangle$.

Proof. Differentiating the ansatz (5.39) for the solution we obtain

$$\begin{aligned}\langle \dot{Y}, X \rangle &= \langle Y_0, -U^{-1} \dot{U} U^{-1} X U + U^{-1} X \dot{U} \rangle \\ &= \langle Y_0, U^{-1} [X, \dot{U} U^{-1}] U \rangle = \langle Y, [X, \dot{U} U^{-1}] \rangle,\end{aligned}$$

where we have used (5.39) in the first and the last equation. This shows that (5.37) is satisfied with the choice $\dot{U} U^{-1} = -H'(Y)$. \square

Example 5.7 (Rigid Body). The Lie group corresponding to the rigid body is $\mathrm{SO}(3)$ with the Lie algebra $\mathfrak{so}(3)$ of skew-symmetric 3×3 matrices, with the basis

$$E_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad E_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad E_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

If we let $x = (x_1, x_2, x_3)^T$ be the coordinates of $X \in \mathfrak{so}(3)$, then we have $Xv = x \times v$ for all $v \in \mathbb{R}^3$. Since for $U \in \mathrm{SO}(3)$,

$$U^{-1} X U v = U^{-1} (x \times U v) = U^{-1} x \times v,$$

the vector $U^{-1} x$ consists of the coordinates of $U^{-1} X U \in \mathfrak{so}(3)$.

Let $y = (y_1, y_2, y_3)^T$ be the coordinates of $Y \in \mathfrak{so}(3)^*$ with respect to the dual basis of E_1, E_2, E_3 . Since

$$\langle Y, U^{-1} X U \rangle = \left\langle \sum_{j=1}^3 y_j F_j, \sum_{i=1}^3 (U^{-1} x)_i E_i \right\rangle = y^T U^{-1} x = (U y)^T x,$$

the coordinates of $\mathrm{Ad}_{U^{-1}} Y$ are given by the vector $U y$. Therefore, the coordinates of the coadjoint orbit of Y lie on a sphere of radius $\|y\|$. The conservation of the coadjoint orbit thus reduces here to the preservation of the Casimir $C(y) = y_1^2 + y_2^2 + y_3^2$.

Lie–Poisson integrators seem to have first been considered by Ge & Marsden (1988), who extend the construction of symplectic methods by generating functions to Lie–Poisson systems. Channel & Scovel (1991) propose an implementation of these methods based on a coordinatization of the group by the exponential map.

McLachlan (1993) proposes integrators based on splitting the Hamiltonian and illustrates this approach for various examples of Lie–Poisson systems. When applicable, such splitting integrators yield Poisson integrators that preserve the coadjoint orbits, since they are composed of exact flows of Lie–Poisson systems.

Engø & Faltinsen (2001) propose to solve numerically the Lie–Poisson system (5.34) by applying Lie group integrators such as those of Sect. IV.8 to the differential equation (5.40) with (5.39). This approach keeps the solution on the coadjoint orbit by construction, but it does not, in general, give a Poisson integrator.

VII.5.5 Lie–Poisson Reduction

The reduction of the Hamiltonian equations of motion of the free rigid body to the Euler equations is an instance of a general reduction process from Hamiltonian systems with symmetry on a Lie group to Lie–Poisson systems, which we now describe; cf. Marsden & Ratiu (1999), Chap. 13, for a presentation in a more abstract framework and for an historical account.

Let us assume that the Lie group G is a subgroup of $GL(n)$ given by

$$G = \{Q; g_i(Q) = 0, i = 1, \dots, m\}, \quad (5.42)$$

and consider a Hamiltonian system on G ,

$$\begin{aligned} \dot{P} &= -\nabla_Q H(P, Q) - \sum_{i=1}^m \lambda_i \nabla_Q g_i(Q), & \dot{Q} &= \nabla_P H(P, Q) \\ 0 &= g_i(Q), & i &= 1, \dots, m, \end{aligned} \quad (5.43)$$

where P, Q are square matrices, and $\nabla_Q H = (\partial H / \partial Q_{ij})$. This is of the form discussed in Sect. VII.1.2. In regions where the matrix

$$\left(\frac{\partial^2 H(P, Q)}{\partial P^2} \left(\nabla_Q g_i(Q), \nabla_Q g_j(Q) \right) \right)_{i,j=1}^m \quad \text{is invertible,} \quad (5.44)$$

the Lagrange parameters λ_i can be expressed in terms of P and Q (cf. formula (1.13)). Hence, a unique solution exists locally provided the initial values (P_0, Q_0) are consistent, i.e., $g_i(Q_0) = 0$ and

$$g'_i(Q_0) \left(\nabla_P H(P_0, Q_0) \right) = \text{trace} \left(\nabla_Q g_i(Q_0)^T \nabla_P H(P_0, Q_0) \right) = 0,$$

or equivalently, $Q_0 \in G$ and $\nabla_P H(P_0, Q_0) \in T_{Q_0} G$.

We now assume that the Hamiltonian H is quadratic in P . As we have seen in Sect. VII.1.2, the equations (5.43) can be viewed as a differential equation on the cotangent bundle $T^*G = \{(P, Q); Q \in G, P \in T_Q^*G\}$, where the cotangent space T_Q^*G is identified with a subspace of matrices such that

$$P \in T_Q^*G \quad \text{if and only if} \quad \nabla_P H(P, Q) \in T_Q G. \quad (5.45)$$

With this identification, the duality between T_Q^*G and $T_Q G$ is given by the matrix inner product

$$\langle P, V \rangle = \text{trace}(P^T V) \quad \text{for } P \in T_Q^*G, V \in T_Q G.$$

We call the Hamiltonian *left-invariant*, if

$$H(U^T P, U^{-1} Q) = H(P, Q) \quad \text{for all } U \in G. \quad (5.46)$$

In this case we have $H(P, Q) = H(Q^T P, I)$ and by differentiating we obtain $\nabla_P H(P, Q) = Q \nabla_P H(Q^T P, I)$. By (5.45) and since $T_Q G = \{QX; X \in \mathfrak{g}\}$ with the Lie algebra $\mathfrak{g} = T_I G$ (cf. Sect. IV.6), this relation implies

$$P \in T_Q^* G \quad \text{if and only if} \quad Q^T P \in T_I^* G = \mathfrak{g}^*. \quad (5.47)$$

Now $H(P, Q)$ depends, for $(P, Q) \in T^* G$, only on the product $Y = Q^T P \in \mathfrak{g}^*$, and we write⁷ $H(P, Q) = H(Y)$ with a function $H : \mathfrak{g}^* \rightarrow \mathbb{R}$.

Left-invariant Hamiltonian systems can be reduced to a Lie–Poisson system of half the dimension by a process that generalizes the derivation of the Euler equations for the rigid body.

Theorem 5.8. *Consider a Hamiltonian system (5.43) on a matrix Lie group G with a left-invariant quadratic Hamiltonian $H(P, Q) = H(Y)$ for $Y = Q^T P$. If $(P(t), Q(t)) \in T^* G$ is a solution of the system (5.43), then $Y(t) = Q(t)^T P(t) \in \mathfrak{g}^*$ solves the differential equation (5.37).*

Proof. It is convenient for the proof (though not necessary, see the lines following (2.17)) to extend the Hamiltonian $H : \mathfrak{g}^* \rightarrow \mathbb{R}$ to a function of arbitrary matrices Y by setting $H(Y) = H(\Pi Y)$, where Π is the projection onto \mathfrak{g}^* given by $\langle \Pi Y, X \rangle = (Y, X)$ for all $X \in \mathfrak{g}$, with the matrix inner product $(Y, X) = \text{trace}(Y^T X)$.

We first compute the derivatives of $H(P, Q) = H(Y) = H(\Pi Y) = H(y)$ where $Q^T P = Y$ and, using the notation of Theorem 5.5, $\Pi Y = \sum_{j=1}^d y_j F_j$. Since $y_j = \langle \Pi Q^T P, E_j \rangle = (Q^T P, E_j)$, it follows from $\nabla_A \text{trace}(A^T B) = B$ that

$$\nabla_P H(P, Q) = \sum_{j=1}^d \frac{\partial H(y)}{\partial y_j} \nabla_P y_j = \sum_{j=1}^d \frac{\partial H(y)}{\partial y_j} \nabla_P \text{trace}(P^T Q E_j) = Q H'(Y), \quad (5.48)$$

where $H'(Y) = \sum_{j=1}^d \frac{\partial H(y)}{\partial y_j} E_j \in \mathfrak{g}$ as in Theorem 5.5. Using the identity $y_j = \text{trace}(P^T Q E_j) = \text{trace}(Q^T P E_j^T)$ we get in a similar way

$$\nabla_Q H(P, Q) = P H'(Y)^T. \quad (5.49)$$

Consequently, the differential equations (5.43) become

$$\dot{P} = -P H'(Q^T P)^T - \sum_{i=1}^m \lambda_i \nabla_Q g_i(Q), \quad \dot{Q} = Q H'(Q^T P). \quad (5.50)$$

The product rule $\dot{Y} = \dot{Q}^T P + Q^T \dot{P}$ for $Y = Q^T P$ thus yields

$$\dot{Y} = H'(Y)^T Y - Y H'(Y)^T - \sum_{i=1}^m \lambda_i Q^T \nabla_Q g_i(Q). \quad (5.51)$$

⁷ We use again the same letter for different functions. Since they have either one or two arguments, no confusion should arise.

For $X \in \mathfrak{g}$, we now exploit the properties

$$\begin{aligned}\langle Q^T \nabla_Q g_i(Q), X \rangle &= \langle \nabla_Q g_i(Q), QX \rangle = 0 \quad (\text{because } QX \in T_Q G) \\ \langle [H'(Y)^T, Y], X \rangle &= \text{trace}((Y^T H'(Y) - H'(Y) Y^T) X) \\ &= \text{trace}(Y^T (H'(Y) X - X H'(Y))) = \langle Y, [H'(Y), X] \rangle.\end{aligned}$$

Since $Y(t) \in \mathfrak{g}^*$ for all t , this gives the equation (5.37). \square

Reduced System and Reconstruction. Combining Theorems 5.8 and 5.5, we have reduced the Hamiltonian system (5.43) to the Lie–Poisson system for $y(t) \in \mathbb{R}^d$,

$$\dot{y} = B(y) \nabla H(y), \quad (5.52)$$

of half the dimension. To recover the full solution $(P(t), Q(t)) \in T^*G$, we must solve this system along with

$$\dot{Q} = QH'(Y), \quad P = Q^{-T}Y \quad (5.53)$$

where $Y = \sum_{j=1}^d y_j F_j \in \mathfrak{g}^*$.

Poisson Structures. The Poisson bracket on \mathbb{R}^d defined by $B(y)$ is still closely related to the canonical Poisson bracket on \mathbb{R}^{2n^2} . Consider left-invariant real-valued functions K, L on T^*G . These can be viewed as functions on $T^*G/G = \mathfrak{g}^* \subset \mathbb{R}^{n \times n}$,

$$K(P, Q) = K(Y) \quad \text{for } Y = Q^T P.$$

(As we did previously in this section, we use the same symbol for these functions.) Via the projection $\Pi : \mathbb{R}^{n \times n} \rightarrow \mathfrak{g}^*$ used in the above proof, we can extend $K(Y) = K(\Pi Y)$ to arbitrary $n \times n$ matrices Y , and via the above relation to a left-invariant function $K(P, Q)$ on $\mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n}$, on which we have the canonical Poisson bracket

$$\{K, L\}_{\text{can}} = \sum_{k,l=1}^n \left(\frac{\partial K}{\partial Q_{kl}} \frac{\partial L}{\partial P_{kl}} - \frac{\partial K}{\partial P_{kl}} \frac{\partial L}{\partial Q_{kl}} \right).$$

On the other hand, we can view K as a function on \mathbb{R}^d by choosing coordinates on \mathfrak{g}^* ,

$$K(y) = K(Y) \quad \text{for } Y = \sum_{j=1}^d y_j F_j \in \mathfrak{g}^*.$$

On \mathbb{R}^d we have the Poisson bracket defined by the structure matrix $B(y)$,

$$\{K, L\} = \sum_{i,j=1}^d \frac{\partial K}{\partial y_i} b_{ij} \frac{\partial L}{\partial y_j}.$$

Lemma 5.9. *For left-invariant functions K, L as described above, we have for $Q^T P = Y = \sum_{j=1}^d y_j F_j \in \mathfrak{g}^*$*

$$\{K, L\}(y) = \langle Y, [L'(Y), K'(Y)] \rangle = \{K, L\}_{\text{can}}(P, Q)$$

where $K'(Y) = \sum_{i=1}^d \frac{\partial K}{\partial y_i}(y) E_i \in \mathfrak{g}$.

Proof. The first equality follows from the identity

$$b_{ij}(y) = \langle Y, [E_j, E_i] \rangle,$$

which is a direct consequence of the definition (5.35) with (5.36). For the second equality, the relations (5.48) and (5.49) for K and L yield

$$\begin{aligned} \{K, L\}_{\text{can}}(P, Q) &= \text{trace} (K'(Y) Y^T L'(Y) - K'(Y)^T Y L'(Y)^T) \\ &= \text{trace} (K'(Y) Y^T L'(Y) - L'(Y) Y^T K'(Y)) \\ &= \text{trace} (Y^T [L'(Y), K'(Y)]) = \langle Y, [L'(Y), K'(Y)] \rangle, \end{aligned}$$

which is the result. \square

Discrete Lie–Poisson Reduction. Consider a symplectic integrator

$$(P_1, Q_1) = \Phi_h(P_0, Q_0) \quad \text{on } T^*G$$

for the left-invariant Hamiltonian system (5.43), and suppose that the method preserves the left-invariance: if $\Phi_h(P_0, Q_0) = (P_1, Q_1)$, then

$$\Phi_h(U^T P_0, U^{-1} Q_0) = (U^T P_1, U^{-1} Q_1) \quad \text{for all } U \in G. \quad (5.54)$$

For example, this is satisfied by the RATTLE algorithm. The method then induces a one-step map

$$Y_1 = \Psi_h(Y_0) \quad \text{on } \mathfrak{g}^*$$

by setting $Y_1 = Q_1^T P_1$ for $(P_1, Q_1) = \Phi_h(P_0, Q_0)$ with $Q_0^T P_0 = Y_0$. This is a numerical integrator for (5.37), and in the coordinates $y = (y_j)$ with respect to the basis (F_j) of \mathfrak{g}^* this gives a map

$$y_1 = \psi_h(y_0) \quad \text{on } \mathbb{R}^d,$$

which is a numerical integrator for the Poisson system (5.52).

Example 5.10. For the rigid body, applying the RATTLE algorithm to the constrained Hamiltonian system (5.18) yields the integrator for the Euler equations discussed in Sect. VII.5.3. By the following result this is a Poisson integrator.

Theorem 5.11. *If $\Phi_h(P, Q)$ is a symplectic and left-invariant integrator for (5.43), then its reduction $\psi_h(y)$ is a Poisson map.*

Proof. We write ψ_h as the composition

$$\psi_h : \mathbb{R}^d \xrightarrow{\xi} T^*G \xrightarrow{\Phi_h} T^*G \xrightarrow{\eta} \mathbb{R}^d$$

where $\eta = (\eta_j)$ is the function with $\eta_j(P, Q) = y_j$ for $Q^T P = \sum_{j=1}^d y_j F_j$, and ξ is any right inverse of η , i.e., $\eta \circ \xi = \text{id}$. For arbitrary smooth real-valued functions K, L on \mathbb{R}^d we then have for $(P, Q) = \xi(y)$, using Lemma 5.9 in the outer equalities and the symplecticity of Φ_h in the middle equality,

$$\begin{aligned} \{K \circ \psi_h, L \circ \psi_h\}(y) &= \{K \circ \eta \circ \Phi_h, L \circ \eta \circ \Phi_h\}_{\text{can}}(P, Q) \\ &= \{K \circ \eta, L \circ \eta\}_{\text{can}}(\Phi_h(P, Q)) = \{K, L\}(\psi_h(y)). \end{aligned}$$

This equation states that ψ_h is a Poisson map. \square

A similar reduction in a discrete Lagrangian framework is studied by Marsden, Pekarsky & Shkoller (1999).

The reduced numerical maps ψ_h and Ψ_h have further structure-preserving properties: they preserve the Casimirs and the co-adjoint orbits. This will be shown in Sect. IX.5.3 with the help of backward error analysis.

VII.6 Reduced Models of Quantum Dynamics

To incorporate quantum effects in molecular dynamics simulations, computations are done with models that are intermediate between classical molecular dynamics based on Newton's equations of motion and full quantum dynamics described by the N -particle Schrödinger equation. The direct computational treatment of the latter is not feasible because of its high dimensionality. These intermediate models are obtained by the Hamiltonian reduction (2.17) from an infinite-dimensional Hilbert space to an appropriately chosen manifold. In chemical physics, this reduction is known as the *Dirac–Frenkel time-dependent variational principle*. We illustrate this procedure for the case where the quantum-mechanical wave function is approximated by a complex Gaussian as proposed by Heller (1975). It turns out that the resulting ordinary differential equations have a Poisson structure, which was recently described by Faou & Lubich (2004). Following that paper, we derive a structure-preserving explicit integrator for Gaussian wavepackets, which tends to the Störmer–Verlet method in the classical limit.

VII.6.1 Hamiltonian Structure of the Schrödinger Equation

The introduction of wave mechanics stands ... as Schrödinger's monument and a worth one. (From Schrödinger's obituary in *The Times* 1961; quoted from <http://www-groups.dcs.st-and.ac.uk/history/Mathematicians/Schrodinger.html>)

The time-dependent N -body Schrödinger equation reads (see, e.g., Messiah (1999) and Thaller (2000))

$$i\varepsilon \frac{\partial \psi}{\partial t} = H\psi \quad (6.1)$$

for the wave function $\psi = \psi(x, t)$ depending on the spatial variables $x = (x_1, \dots, x_N)$ with $x_k \in \mathbb{R}^d$ (e.g., with $d = 1$ or 3 in the partition) and the time $t \in \mathbb{R}$. The squared absolute value $|\psi(x, t)|^2$ represents the joint probability density for N particles to be at the positions x_1, \dots, x_N at time t . In (6.1), ε is a (small) positive number representing the scaled Planck constant and i is the complex imaginary unit. The Hamiltonian operator H is written

$$H = T + V$$

with the kinetic and potential energy operators

$$T = - \sum_{k=1}^N \frac{\varepsilon^2}{2m_k} \Delta_{x_k} \quad \text{and} \quad V = V(x),$$

where $m_k > 0$ is a particle mass and Δ_{x_k} the Laplacian in the variable $x_k \in \mathbb{R}^d$, and where the real-valued potential V acts as a multiplication operator $(V\phi)(x) = V(x)\phi(x)$. Under appropriate conditions on V (boundedness of V is sufficient, but by no means necessary), the operator H is then a self-adjoint operator on the complex Hilbert space $L^2(\mathbb{R}^{dN}, \mathbb{C})$ with domain $D(H) = D(T) = \{\phi \in L^2(\mathbb{R}^{dN}, \mathbb{C}); T\phi \in L^2(\mathbb{R}^{dN}, \mathbb{C})\}$; see Sect. V.5.3 of Kato (1980).

We separate the real and imaginary parts of $\psi = v + iw \in L^2(\mathbb{R}^{dN}, \mathbb{C})$, the complex Hilbert space of Lebesgue square-integrable functions. The functions v and w are thus functions in the *real* Hilbert space $L^2(\mathbb{R}^{dN}, \mathbb{R})$. We denote the complex inner product by $\langle \cdot, \cdot \rangle$ and the real inner product by (\cdot, \cdot) . The L^2 norms will be simply denoted by $\|\cdot\|$.

As H is a real operator, formula (6.1) can be written

$$\begin{aligned} \varepsilon \dot{v} &= Hw, \\ \varepsilon \dot{w} &= -Hv, \end{aligned} \quad (6.2)$$

or equivalently, with the canonical structure matrix

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

and the Hamiltonian function (we use the same symbol H as for the operator)

$$H(v, w) = \frac{1}{2} \langle \psi, H\psi \rangle = \frac{1}{2} (v, Hv) + \frac{1}{2} (w, Hw)$$

for $\psi = v + iw$ in the domain of the operator H . This becomes the canonical Hamiltonian system

$$\begin{pmatrix} \dot{v} \\ \dot{w} \end{pmatrix} = \varepsilon^{-1} J^{-1} \nabla H(v, w).$$

Note that the real multiplication with J corresponds to the complex multiplication with the imaginary unit i . The flow of this system preserves the canonical symplectic two-form

$$\omega(\xi_1, \xi_2) = (J\xi_1, \xi_2), \quad \xi_1, \xi_2 \in L^2(\mathbb{R}^{dN}, \mathbb{R})^2. \quad (6.3)$$

VII.6.2 The Dirac–Frenkel Variational Principle

For dealing with atoms involving many electrons the accurate quantum theory, involving a solution of the wave equation in many-dimensional space, is far too complicated to be practicable. One must therefore resort to approximate methods. (P.A.M. Dirac 1930)

Reduced models of the Schrödinger equation are obtained by restricting the equation to an approximation manifold \mathcal{M} via (2.17), viz.,

$$(\varepsilon J\dot{u} - \nabla H(u), \xi) = 0 \quad \text{for all } \xi \in T_u\mathcal{M}, \quad (6.4)$$

or equivalently in complex notation for $u = (v, w)^T = v + iw$,

$$\operatorname{Re} \langle \varepsilon i\dot{u} - Hu, \xi \rangle = 0 \quad \text{for all } \xi \in T_u\mathcal{M}. \quad (6.5)$$

Taking the real part can be omitted if the tangent space $T_u\mathcal{M}$ is complex linear. Equation (6.5) (usually without the real part) is known as the Dirac–Frenkel time-dependent variational principle, after Dirac (1930) and Frenkel (1934); see also McLachlan (1964), Heller (1976), Beck, Jäckle, Worth & Meyer (2000), and references therein.

We choose a (local) coordinate map $u = \chi(y)$ of \mathcal{M} and denote its derivative $X_{\mathbb{C}}(y) = V(y) + iW(y) = \chi'(y)$ or in the real setting as $X = \begin{pmatrix} V \\ W \end{pmatrix}$. Denoting by X^T the adjoint of X with respect to the real inner product (\cdot, \cdot) , we thus obtain

$$\varepsilon X(y)^T JX(y) \dot{y} = X(y)^T \nabla_u H(\chi(y)).$$

With $X_{\mathbb{C}}^*$ denoting the adjoint of $X_{\mathbb{C}}$ with respect to the complex inner product $\langle \cdot, \cdot \rangle$, we note $X_{\mathbb{C}}^* X_{\mathbb{C}} = (V^T V + W^T W) + i(V^T W - W^T V) = X^T X - iX^T JX$ and hence

$$X^T JX = -\operatorname{Im} X_{\mathbb{C}}^* X_{\mathbb{C}}. \quad (6.6)$$

Lemma 6.1. *If $T_u\mathcal{M}$ is a complex linear space for every $u \in \mathcal{M}$, then \mathcal{M} is a symplectic submanifold of $L^2(\mathbb{R}^N, \mathbb{R})^2$, that is, the symplectic two-form (6.3) is non-degenerate on $T_u\mathcal{M}$ for all $u \in \mathcal{M}$. Expressed in coordinates,*

$$X(y)^T JX(y) \text{ is invertible for all } y.$$

Proof. We fix $u = \chi(y) \in \mathcal{M}$ and omit the argument y in the following. Since $T_u\mathcal{M} = \operatorname{Range}(X_{\mathbb{C}})$ is complex linear by assumption, there exists a real linear mapping $L : \mathbb{R}^m \rightarrow \mathbb{R}^m$ such that $iX_{\mathbb{C}}\eta = X_{\mathbb{C}}L\eta$ for all $\eta \in \mathbb{R}^m$. This implies

$$JX = XL \quad \text{and} \quad L^2 = -\operatorname{Id}$$

and hence $X^T JX = X^T XL$, which is invertible. \square

Approximation properties of the Dirac–Frenkel variational principle can be obtained from the interpretation as the *orthogonal* projection $\dot{u} = P_{\perp}(u) \frac{1}{i\varepsilon} Hu$, which corresponds to taking the imaginary part in (6.5), as opposed to the symplectic projection in (6.4) which corresponds to the real part. See Lubich (2005) for a near-optimality result for approximation on the manifold.

VII.6.3 Gaussian Wavepacket Dynamics

We develop a new approach to semiclassical dynamics which exploits the fact that extended wavefunctions for heavy particles (or particles in harmonic potentials) may be decomposed into time-dependent wave packets, which spread minimally and which execute classical or nearly classical trajectories. A Gaussian form for the wave packets is assumed and equations of motion are derived for the parameters characterizing the Gaussian. (E.J. Heller 1975)

The variational Gaussian wavepacket dynamics of Heller (1976) is obtained by choosing the manifold \mathcal{M} in (6.5) as consisting of complex Gaussians. For ease of presentation we restrict our attention in the following to the one-particle case $N = 1$ (the extension to $N > 1$ is straightforward; cf. Heller (1976) and Faou & Lubich (2004)). Here we have

$$\mathcal{M} = \{u = \chi(y) \in L^2(\mathbb{R}^d, \mathbb{C}) : y = (p, q, \alpha, \beta, \gamma, \delta) \in \mathbb{R}^{2d+4} \text{ with } \beta > 0\} \quad (6.7)$$

with

$$\left(\chi(y)\right)(x) = \exp\left(\frac{i}{\varepsilon}\left((\alpha + i\beta)|x - q|^2 + p \cdot (x - q) + \gamma + i\delta\right)\right), \quad (6.8)$$

where $|\cdot|$ and \cdot stand for the Euclidean norm and inner product on \mathbb{R}^d . The parameters q and p represent the average position and momentum, respectively: for $u = \chi(y)$ with $y = (p, q, \alpha, \beta, \gamma, \delta)$ and $\|u\| = 1$, a direct calculation shows that

$$q = \langle u, xu \rangle = \int_{\mathbb{R}^d} x |u(x)|^2 dx, \quad p = \langle u, -i\varepsilon \nabla u \rangle.$$

The parameter $\beta > 0$ determines the width of the wavepacket. The tangent space $T_u \mathcal{M} \subset L^2(\mathbb{R}^d, \mathbb{C})$ at a given point $u = \chi(y) \in \mathcal{M}$ is $(2d + 4)$ -dimensional and is made of the elements of $L^2(\mathbb{R}^d, \mathbb{C})$ written as

$$\frac{i}{\varepsilon} \left((A + iB)|x - q|^2 + (P - 2(\alpha + i\beta)Q) \cdot (x - q) - p \cdot Q + C + iD \right) u \quad (6.9)$$

with arbitrary $(P, Q, A, B, C, D)^T \in \mathbb{R}^{2d+4}$. We note that $T_u \mathcal{M}$ is complex linear, and $u \in T_u \mathcal{M}$. By choosing $\xi = iu$ in (6.5), this yields $(d/dt)\|u\|^2 = 2 \operatorname{Re} \langle \dot{u}, u \rangle = 0$ and hence the preservation of the squared L^2 norm of $u = \chi(y)$, which is given by

$$\begin{aligned}
I(y) &= \|u\|^2 = \int_{\mathbb{R}^d} |u(x)|^2 dx \\
&= \int_{\mathbb{R}^d} \exp\left(-\frac{2}{\varepsilon}(\beta|x-q|^2 + \delta)\right) dx = \exp\left(-\frac{2\delta}{\varepsilon}\right) \left(\frac{\pi\varepsilon}{2\beta}\right)^{d/2}.
\end{aligned} \tag{6.10}$$

The physically reasonable situation is $\|u\|^2 = 1$, which corresponds to the interpretation of $|u(x)|^2$ as a probability density.

With these preparations we have the following result of Faou & Lubich (2004).

Theorem 6.2. *The Hamiltonian reduction of the Schrödinger equation to the Gaussian wavepacket manifold \mathcal{M} of (6.7)-(6.8) yields the Poisson system*

$$\dot{y} = B(y)\nabla K(y) \tag{6.11}$$

where, for $y = (p, q, \alpha, \beta, \gamma, \delta) \in \mathbb{R}^{2d+4}$ with $\beta > 0$, and with 1_d denoting the d -dimensional identity,

$$B(y) = \frac{1}{I(y)} \begin{pmatrix} 0 & -1_d & 0 & 0 & -p & 0 \\ 1_d & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{4\beta^2}{\varepsilon d} & 0 & -\beta \\ 0 & 0 & -\frac{4\beta^2}{\varepsilon d} & 0 & \beta & 0 \\ p^T & 0 & 0 & -\beta & 0 & \frac{d+2}{4}\varepsilon \\ 0 & 0 & \beta & 0 & -\frac{d+2}{4}\varepsilon & 0 \end{pmatrix} \tag{6.12}$$

defines a Poisson structure, and for $u = \chi(y)$,

$$K(y) = \langle u, Hu \rangle = K_T(y) + K_V(y) \tag{6.13}$$

is the total energy, with kinetic and potential parts

$$K_T(y) = I(y) \left(\frac{|p|^2}{2m} + \frac{\varepsilon d}{2m} \frac{\alpha^2 + \beta^2}{\beta} \right) = \langle u, Tu \rangle$$

and

$$K_V(y) = \int_{\mathbb{R}^d} V(x) \exp\left(-\frac{2}{\varepsilon}(\beta|x-q|^2 + \delta)\right) dx = \langle u, Vu \rangle.$$

Both $K(y)$ and $I(y)$ are first integrals of the system.

Proof. As in (2.22), the differential equation for y is $\varepsilon X(y)^T J X(y) \dot{y} = \frac{1}{2} \nabla K(y)$. We note (6.6) and

$$X_{\mathbb{C}}(y) = \frac{i}{\varepsilon} (x - q, -2a(x - q) - p, |x - q|^2, i|x - q|^2, 1, i) u$$

where $a = \alpha + i\beta$ and $u = \chi(y)$ in the complex setting. Using the calculus of Gaussian integrals, we compute

$$\varepsilon X^T(y)JX(y) = \frac{1}{2} I(y) \begin{pmatrix} 0 & 1_d & 0 & 0 & 0 & 0 \\ -1_d & 0 & 0 & \frac{dp}{2\beta} & 0 & \frac{2p}{\varepsilon} \\ 0 & 0 & 0 & -\frac{\varepsilon d(d+2)}{8\beta^2} & 0 & -\frac{d}{2\beta} \\ 0 & -\frac{dp^T}{2\beta} & \frac{\varepsilon d(d+2)}{8\beta^2} & 0 & \frac{d}{2\beta} & 0 \\ 0 & 0 & 0 & -\frac{d}{2\beta} & 0 & -\frac{2}{\varepsilon} \\ 0 & -\frac{2p^T}{\varepsilon} & \frac{d}{2\beta} & 0 & \frac{\varepsilon}{2} & 0 \end{pmatrix},$$

and inversion yields the differential equation with $B(y) = (2\varepsilon X^T(y)JX(y))^{-1}$ as stated. The system is a Poisson system by Theorem 2.8. \square

Assuming $I(y) = \|u\|^2 = 1$, we observe that the differential equations for the average position and momentum, q and p , read

$$\dot{q} = p/m, \quad \dot{p} = -\langle u, \nabla V u \rangle \quad (6.14)$$

for $u = \chi(y)$ and $y = (p, q, \alpha, \beta, \gamma, \delta)$. We then note $\langle u, \nabla V u \rangle \rightarrow \nabla V(q)$ as $\varepsilon \rightarrow 0$. The differential equations for q and p thus tend to Newtonian equations of motion in the classical limit $\varepsilon \rightarrow 0$:

$$\dot{q} = p/m, \quad \dot{p} = -\nabla V(q). \quad (6.15)$$

It will be useful to consider also scaled variables

$$\hat{y} = (p, q, \alpha, \hat{\beta}, \gamma, \hat{\delta}) \in \mathbb{R}^{2d+4} \quad \text{with} \quad \hat{\beta} = \frac{\beta}{\varepsilon}, \quad \hat{\delta} = \frac{\delta}{\varepsilon}. \quad (6.16)$$

Here we have

$$\dot{\hat{y}} = \hat{B}(\hat{y}) \nabla \hat{K}(\hat{y}) \quad (6.17)$$

where the structure matrix $\hat{B}(\hat{y})$ is independent of ε , and where $\hat{K}(\hat{y})$ depends regularly on $\varepsilon \geq 0$.

VII.6.4 A Splitting Integrator for Gaussian Wavepackets

With the natural splitting $H = T + V$ into kinetic and potential energy, we now consider the variational splitting integrator (4.7) – (4.8), which here becomes the following.

1. We define u_n^+ in \mathcal{M} as the solution at time $h/2$ of the equation for u ,

$$\langle i\varepsilon \dot{u} - Vu, \xi \rangle = 0 \quad \text{for all } \xi \in T_u \mathcal{M} \quad (6.18)$$

with initial value $u(0) = u_n \in \mathcal{M}$.

2. We define u_{n+1}^- as the solution at time h of

$$\langle i\varepsilon \dot{u} - Tu, \xi \rangle = 0 \quad \text{for all } \xi \in T_u \mathcal{M} \quad (6.19)$$

with initial value $u(0) = u_n^+$.

3. Then u_{n+1} is the solution at time $h/2$ of (6.18) with initial value $u(0) = u_{n+1}^-$.

By Theorem 6.2, the substeps in the definition of this splitting method written in the coordinates $y = (p, q, \alpha, \beta, \gamma, \delta)$ are the exact flows $\varphi_{h/2}^V$ and φ_h^T of the Poisson systems

$$\dot{y} = B(y)\nabla K_V(y) \quad \text{and} \quad \dot{y} = B(y)\nabla K_T(y).$$

Note that both equations preserve the L^2 norm of $u = \chi(y)$, which we assume to be 1 in the following.

Most remarkably, these equations can be solved explicitly. Let us consider first the equations (6.19). They are written, for $a = \alpha + i\beta$ and $c = \gamma + i\delta$, as

$$\begin{cases} \dot{q} = p/m, \\ \dot{p} = 0, \end{cases} \quad \begin{cases} \dot{a} = -2a^2/m, \\ \dot{c} = (\frac{1}{2}|p|^2 + i\varepsilon da)/m, \end{cases} \quad (6.20)$$

with initial values $y_0 = (p_0, q_0, a_0, c_0)$ corresponding to $u_0 = \chi(y_0)$. They have the solution

$$q(t) = q_0 + \frac{t}{m} p_0, \quad p(t) = p_0, \quad a(t) = \frac{a_0}{1 + 2a_0 t/m},$$

and

$$c(t) = c_0 + t \frac{|p_0|^2}{2m} + \frac{i\varepsilon d}{2} \log \left(1 + \frac{2a_0 t}{m} \right).$$

Let us now consider the equations (6.18). Taking into account the fact that the potential V is real, these equations are written

$$\begin{aligned} \dot{p} &= -\langle u, \nabla V u \rangle, & \dot{q} &= 0, \\ \dot{\alpha} &= -\frac{1}{2d} \langle u, \Delta V u \rangle, & \dot{\beta} &= 0, \\ \dot{\gamma} &= -\langle u, V u \rangle + \frac{\varepsilon}{8\beta} \langle u, \Delta V u \rangle, & \dot{\delta} &= 0, \end{aligned} \quad (6.21)$$

with the L^2 inner products

$$\langle u, W u \rangle = \int_{\mathbb{R}^d} W(x) \exp \left(-\frac{2}{\varepsilon} (\beta |x - q|^2 + \delta) \right) dx \quad (6.22)$$

for $W = V, \nabla V, \Delta V$. As the L^2 inner products in the equations for p, α, γ depend only on q, β, δ which are constant along this trajectory, these equations can be solved trivially, requiring only the computation of the inner products at the initial value. We thus see that the splitting scheme $\Phi_h = \varphi_{h/2}^V \circ \varphi_h^T \circ \varphi_{h/2}^V$ can be computed explicitly. This gives the following algorithm (Faou & Lubich 2004).

Algorithm 6.3 (Gaussian Wavepacket Integrator). *A step from time t_n to t_{n+1} , starting from the Gaussian wavepacket $u_n = \chi(p_n, q_n, \alpha_n, \beta_n, \gamma_n, \delta_n)$, proceeds as follows:*

1. With $\langle W \rangle_n = \langle u_n, W u_n \rangle$ given by (6.22) for $W = V, \nabla V, \Delta V$, compute

$$\begin{aligned}
p_{n+1/2} &= p_n - \frac{h}{2} \langle \nabla V \rangle_n \\
\alpha_n^+ &= \alpha_n - \frac{h}{4d} \langle \Delta V \rangle_n \\
\gamma_n^+ &= \gamma_n + \frac{h\varepsilon}{16\beta_n} \langle \Delta V \rangle_n.
\end{aligned} \tag{6.23}$$

2. From the values $p_{n+1/2}$, $a_n^+ = \alpha_n^+ + i\beta_n$ and $c_n^+ = \gamma_n^+ + i\delta_n$ compute q_{n+1} , $a_{n+1}^- = \alpha_{n+1}^- + i\beta_{n+1}$, and $c_{n+1}^- = \gamma_{n+1}^- + i\delta_{n+1}$ via

$$\begin{aligned}
q_{n+1} &= q_n + \frac{h}{m} p_{n+1/2} \\
a_{n+1}^- &= a_n^+ / \left(1 + \frac{2h}{m} a_n^+\right) \\
c_{n+1}^- &= c_n^+ + \frac{i\varepsilon d}{2} \log \left(1 + \frac{2h}{m} a_n^+\right).
\end{aligned} \tag{6.24}$$

3. Compute p_{n+1} , α_{n+1} , γ_{n+1} from

$$\begin{aligned}
p_{n+1} &= p_{n+1/2} - \frac{h}{2} \langle \nabla V \rangle_{n+1} \\
\alpha_{n+1} &= \alpha_{n+1}^- - \frac{h}{4d} \langle \Delta V \rangle_{n+1} \\
\gamma_{n+1} &= \gamma_{n+1}^- + \frac{h\varepsilon}{16\beta_{n+1}} \langle \Delta V \rangle_{n+1}.
\end{aligned} \tag{6.25}$$

Let us collect properties of this algorithm.

Theorem 6.4. *The splitting scheme of Algorithm 6.3 is an explicit, symmetric, second-order numerical method for Gaussian wavepacket dynamics (6.11)–(6.13). It is a Poisson integrator for the structure matrix (6.12), and it preserves the unit L^2 norm of the wavepackets: $\|u_n\| = 1$ for all n .*

In the limit $\varepsilon \rightarrow 0$, the position and momentum approximations q_n , p_n of this method tend to those obtained by applying the Störmer–Verlet method to the associated classical mechanical system (6.15).

The statement for $\varepsilon \rightarrow 0$ follows directly from the equations for $p_{n+1/2}$, q_{n+1} , p_{n+1} and from noting $\langle \nabla V \rangle_n \rightarrow \nabla V(q_n)$.

In view of the small parameter ε , the discussion of the order of the method requires more care. Here it is useful to consider the integrator in the scaled variables $\hat{y} = (p, q, \alpha, \beta/\varepsilon, \gamma, \delta/\varepsilon)$ of (6.16). Since the differential equation (6.17) contains ε only as a regular perturbation parameter, after n steps of the splitting integrator we have the ε -uniform error bound

$$\hat{y}_n - \hat{y}(t_n) = O(h^2),$$

where the constants symbolized by the O -notation are independent of ε and of n and h with $nh \leq \text{Const}$. For the approximation of the absolute values of the Gaussian wavepackets this yields

$$\| |u_n|^2 - |u(t_n)|^2 \| = O(h^2), \quad (6.26)$$

but the approximation of the phases is only such that

$$\|u_n - u(t_n)\| = O(h^2/\varepsilon). \quad (6.27)$$

We refer to Faou & Lubich (2004) for the formulation of the corresponding algorithm for $N > 1$ particles, for further properties such as the exact conservation of linear and angular momentum and the long-time near-conservation of the total energy $\langle u_n, H u_n \rangle$, and for numerical experiments.

VII.7 Exercises

1. Prove that the Poisson bracket (2.8) satisfies the Jacobi identity (2.4) for all functions F, G, H , if and only if it satisfies (2.4) for the coordinate functions y_i, y_j, y_k .

Hint (F. Engel, in Lie's *Gesammelte Abh.* vol. 5, p. 753). If the Jacobi identity is written as in (3.3), we see that there are no second partial derivatives of F (the left hand side is a Lie bracket, the right-hand side has no second derivatives of F anyway). Other permutations show the same result for G and H .

2. For x in an open subset of \mathbb{R}^m , let $A(x) = (a_{ij}(x))$ be an invertible skew-symmetric $m \times m$ -matrix, with

$$\frac{\partial a_{ij}}{\partial x_k} + \frac{\partial a_{ki}}{\partial x_j} + \frac{\partial a_{jk}}{\partial x_i} = 0 \quad \text{for all } i, j, k. \quad (7.1)$$

(a) Show that $B(x) = A(x)^{-1}$ satisfies (2.10) and hence defines a Poisson bracket.

(b) Generalize Theorem 2.8 to Hamiltonian equations (2.18) with the two-form $\omega_x(\xi_1, \xi_2) = \xi_1^T A(x) \xi_2$.

Remark. Condition (7.1) says that ω is a closed differential form.

3. Solve the following first order partial differential equation:

$$3 \frac{\partial F}{\partial y_1} + 2 \frac{\partial F}{\partial y_2} - 5 \frac{\partial F}{\partial y_3} = 0.$$

Result. $f(2y_1 - 3y_2, 5y_2 + 2y_3)$.

4. Find two solutions of the homogeneous system

$$3 \frac{\partial F}{\partial y_1} + \frac{\partial F}{\partial y_2} - 2 \frac{\partial F}{\partial y_3} - 5 \frac{\partial F}{\partial y_4} = 0, \quad 2 \frac{\partial F}{\partial y_1} - \frac{\partial F}{\partial y_2} - 3 \frac{\partial F}{\partial y_4} = 0,$$

such that their gradients are linearly independent.

5. Consider a Poisson system $\dot{y} = B(y) \nabla H(y)$ and a change of coordinates $z = \vartheta(y)$. Prove that in the new coordinates the system is of the form $\dot{z} = \tilde{B}(z) \nabla K(z)$, where $\tilde{B}(z) = \vartheta'(y) B(y) \vartheta'(y)^T$ (cf. formula (3.12)) and $K(z) = H(y)$.

6. Give an elementary proof of Theorem 4.3.

Hint. Define $\delta(t) := \varphi'_t(y)B(y)\varphi'_t(y)^T - B(\varphi_t(y))$. Using the variational equation for (4.1) prove that $\delta(t)$ is the solution of a homogeneous linear differential equation. Therefore, $\delta(0) = 0$ implies $\delta(t) = 0$ for all t .

7. Let $z = \vartheta(y)$ be a transformation taking the Poisson system $\dot{y} = B(y)\nabla H(y)$ to $\dot{z} = \tilde{B}(z)\nabla K(z)$. Prove that $\Phi_h(y)$ is a Poisson integrator for $B(y)$ if and only if $\Psi_h(z) = \vartheta \circ \Phi_h \circ \vartheta^{-1}(z)$ is a Poisson integrator for $\tilde{B}(z)$.
8. Let B be a skew-symmetric but otherwise arbitrary constant matrix, and consider the Poisson system $\dot{y} = B\nabla H(y)$. Prove that every symplectic Runge–Kutta method is a Poisson integrator for such a system.

Hint. Transform B to block-diagonal form.

9. (M.J. Gander 1994). Consider the Lotka–Volterra equation (2.13) with separable Hamiltonian $H(u, v) = K(u) + L(v)$. Prove that

$$u_{n+1} = u_n + hu_nv_nH_v(u_n, v_n), \quad v_{n+1} = v_n - hu_{n+1}v_nH_u(u_{n+1}, v_n)$$

is a Poisson integrator for this system.

10. Find a change of coordinates that transforms the Lotka–Volterra system (2.14) into a Hamiltonian system (in canonical form). Following the approach of Example 4.11 construct Poisson integrators for this system.
11. Prove that the matrix $B(y)$ of Example 2.7 defines a Poisson bracket, by showing that the bracket is given as Dirac's bracket (Dirac 1950)

$$\{F, G\} = \{\hat{F}, \hat{G}\} - \sum_{i,j} \{\hat{F}, c_i\} \gamma_{ij} \{c_j, \hat{G}\}. \quad (7.2)$$

Here F and G are functions of y , \hat{F} and \hat{G} are smooth functions of x satisfying $\hat{F}(\chi(y)) = F(y)$ and $\hat{G}(\chi(y)) = G(y)$, $c_i(x)$ are the constraint functions defining the manifold \mathcal{M} , and γ_{ij} are the entries of the inverse of the matrix $(\{c_i, c_j\})$. The Poisson bracket to the left in (7.2) corresponds to $B(y)$ and those to the right are the canonical brackets evaluated at $x = \chi(y)$. Replacing $\hat{F}(x)$ by $\hat{F}(x) + \sum_k \mu_k(x)c_k(x)$ with $\mu_k(x)$ such that $\{\hat{F}, c_k\} = 0$ on \mathcal{M} eliminates the sum in (7.2) and proves the Jacobi identity for $B(y)$.

Chapter VIII.

Structure-Preserving Implementation

This chapter is devoted to practical aspects of an implementation of geometric integrators. We explain strategies for changing the step size which do not deteriorate the correct qualitative behaviour of the solution. We study multiple time stepping strategies, the effect of round-off in long-time integrations, and the efficient solution of nonlinear systems arising in implicit integration schemes.

VIII.1 Dangers of Using Standard Step Size Control

Another possible shortcoming of the method concerns its behavior when used with a variable step size . . . The integrator completely loses its desirable qualities . . . This can be understood at least qualitatively by realizing that by changing the time step one is in essence continually changing the nearby Hamiltonian . . . (B. Gladman, M. Duncan & J. Candy 1991)

In the previous chapters we have studied symmetric and symplectic integrators, and we have seen an enormous progress in long-time integrations of various problems. Decades ago, a similar enormous progress was the introduction of algorithms with automatic step size control. Naively, one would expect that the blind combination of both techniques leads to even better performances. We shall see by a numerical experiment that this is not the case, a phenomenon observed by Gladman, Duncan & Candy (1991) and Calvo & Sanz-Serna (1992).

We study the long-time behaviour of symplectic methods combined with the following standard step size selection strategy (see e.g., Hairer, Nørsett & Wanner (1993), Sect. II.4). We assume that an expression err_n related to the local error is available for the current step computed with step size h_n (usually obtained with an embedded method). Based on an asymptotic formula $err_n \approx Ch_n^r$ (for $h_n \rightarrow 0$) and on the requirement to get an error close to a user supplied tolerance Tol , we predict a new step size by

$$h_{new} = 0.85 \cdot h_n \left(\frac{Tol}{err_n} \right)^{1/r}, \quad (1.1)$$

where a safety factor 0.85 is included. We then apply the method with step size $h_{n+1} = h_{new}$. If for the new step $err_{n+1} \leq Tol$, the step is accepted and the integration is continued. If $err_{n+1} > Tol$, it is rejected and recomputed with the step size h_{new} obtained from (1.1) with $n + 1$ instead of n . Similar step size strategies are implemented in most codes for solving ordinary differential equations.

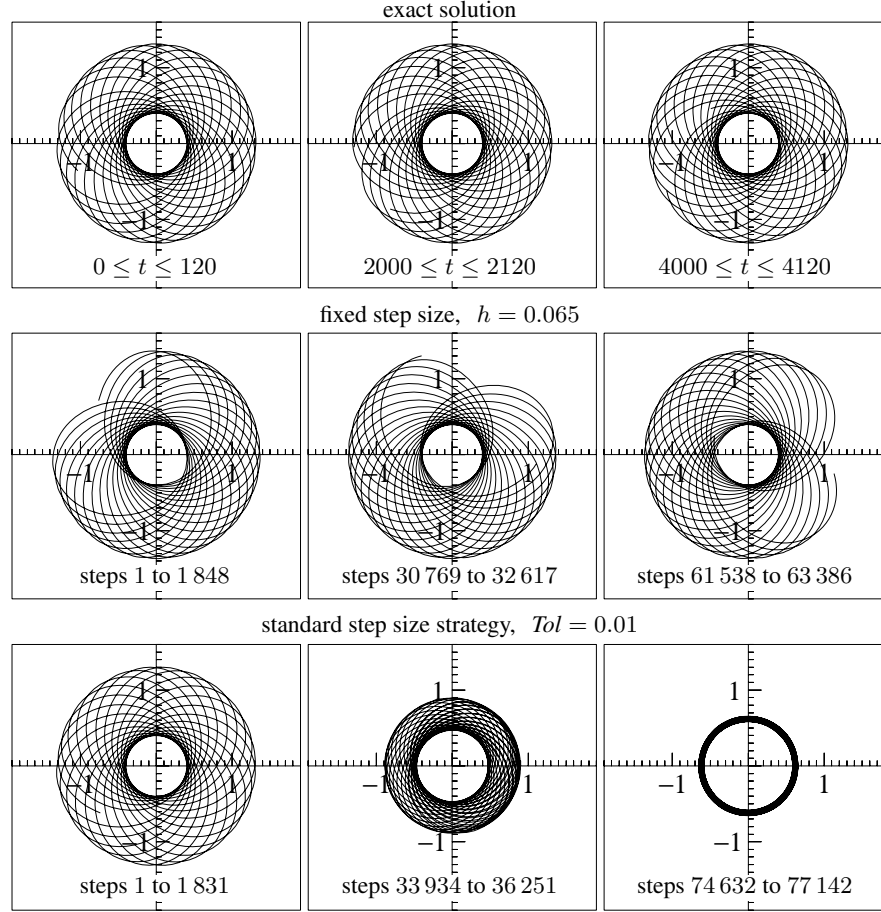


Fig. 1.1. Störmer–Verlet scheme applied with fixed step size (middle) or with the standard step size strategy (below) compared to the exact solution (above); solutions are for the interval $0 \leq t \leq 120$ (left), for $2000 \leq t \leq 2120$ (middle), and for $4000 \leq t \leq 4120$ (right)

Numerical Experiment. We consider the perturbed Kepler problem

$$\begin{aligned} \dot{q}_1 &= p_1, & \dot{p}_1 &= -\frac{q_1}{(q_1^2 + q_2^2)^{3/2}} - \frac{\delta q_1}{(q_1^2 + q_2^2)^{5/2}} \\ \dot{q}_2 &= p_2, & \dot{p}_2 &= -\frac{q_2}{(q_1^2 + q_2^2)^{3/2}} - \frac{\delta q_2}{(q_1^2 + q_2^2)^{5/2}} \end{aligned} \quad (1.2)$$

($\delta = 0.015$) with initial values

$$q_1(0) = 1 - e, \quad q_2(0) = 0, \quad p_1(0) = 0, \quad p_2(0) = \sqrt{(1+e)/(1-e)}$$

(eccentricity $e = 0.6$). As a numerical method we take the *Störmer–Verlet scheme* (I.1.17) which is symmetric, symplectic, and of order 2. The fixed step size imple-

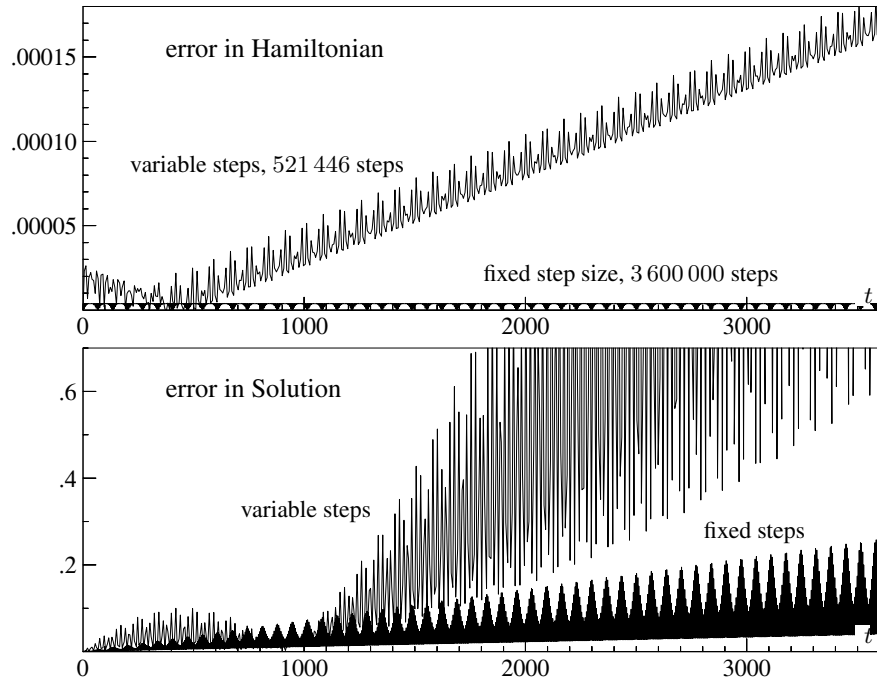


Fig. 1.2. Study of the error in the Hamiltonian and of the global error for the Störmer-Verlet scheme. Fixed step size implementation with $h = 10^{-3}$, variable step size with $Tol = 10^{-4}$

mentation is straightforward. For the variable step size strategy we take for err_n the Euclidean norm of the difference between the Störmer-Verlet solution and the symplectic Euler solution (which is available without any further function evaluation). Since $err_n = \mathcal{O}(h_n^2)$, we take $r = 2$ in (1.1).

The numerical solution in the (q_1, q_2) -plane is presented in Fig. 1.1. To make the long-time behaviour of the two implementations visible, we show the numerical solution on three different parts of the integration interval. We have included the numbers of steps needed for the integration to reach $t = 120$, 2120 , and 4120 , respectively. We see that the qualitative behaviour of the variable step size implementation is not correct, although it is more precise on short intervals. Moreover, the near-preservation of the Hamiltonian is lost (see Fig. 1.2) as is the linear error growth. Apparently, the error in the Hamiltonian behaves like $|a - bt|$ for the variable step size implementation, and that for the solution like $|ct - dt^2|$ (with constants a, b, c, d depending on Tol). Due to the relatively large eccentricity of the problem, the variable step size implementation needs fewer function evaluations for a given accuracy on a short time interval, but the opposite is true for long-time integrations.

The aim of the next two sections is to present approaches which permit the use of variable step sizes for symmetric or symplectic methods without losing the qualitatively correct long-time behaviour.

VIII.2 Time Transformations

A variable step size implementation produces approximations y_n on a (non-equidistant) grid $\{t_n\}$. The same effect can be achieved by performing in advance a time transformation $t \leftrightarrow \tau$ and by applying a constant step size implementation to the transformed system. If the time transformation is given as the solution of a differential equation, it follows from the chain rule $\frac{dy}{d\tau} = \frac{dy}{dt} \frac{dt}{d\tau}$ that the transformed system is

$$y' = \sigma(y)f(y), \quad t' = \sigma(y). \quad (2.1)$$

Here, prime indicates a derivative with respect to τ , and we use the same letter y for the solutions $y(t)$ of $\dot{y} = f(y)$ and $y(\tau)$ of (2.1). If $\sigma(y) > 0$, the correspondence $t \leftrightarrow \tau$ is bijective.

Applying a numerical method with constant step size ε to (2.1) yields approximations $y_n \approx y(\tau_n) = y(t_n)$, where $\tau_n = n\varepsilon$ and

$$t_{n+1} - t_n = \int_{n\varepsilon}^{(n+1)\varepsilon} \sigma(y(\tau)) d\tau \approx \varepsilon \sigma(y_n). \quad (2.2)$$

Approximations to t_n are obtained by integrating numerically the differential equation $t' = \sigma(y)$ together with $y' = \sigma(y)f(y)$.

In the context of geometric numerical integration, we are interested in time transformations such that the vector field $\sigma(y)f(y)$ retains geometric features of $f(y)$.

VIII.2.1 Symplectic Integration

For a Hamiltonian system $\dot{y} = f(y) = J^{-1}\nabla H(y)$ it is natural to search for step size functions $\sigma(y)$ such that (2.1) is again Hamiltonian. For this we have to check whether the Jacobian of $\sigma(y)\nabla H(y)$ is symmetric (cf. Integrability Lemma VI.2.7). But this is the case only if $\nabla H(y)\nabla\sigma(y)^T$ is symmetric, i.e., $\nabla H(y)$ and $\nabla\sigma(y)$ are collinear, so that $\frac{d}{dt}\sigma(y(t)) = \nabla\sigma(y(t))^T J \nabla H(y(t)) = 0$. Consequently, $\sigma(y) = \text{Const}$ along solutions of the Hamiltonian system which is what makes this approach unattractive for a variable step size integration. This disappointing fact has been observed by Stoffer (1988, 1995) and Skeel & Gear (1992).

The main idea for circumventing this difficulty is the following: suppose we want to integrate the Hamiltonian system with steps of size $h \approx \varepsilon \sigma(y)$, where $\sigma(y) > 0$ is a state-dependent given function and $\varepsilon > 0$ is a small parameter. Instead of multiplying the vector field $f(y) = J^{-1}\nabla H(y)$ by $\sigma(y)$, we consider the *new Hamiltonian*

$$K(y) = \sigma(y)(H(y) - H_0), \quad (2.3)$$

where $H_0 = H(y_0)$ for a fixed initial value y_0 . The corresponding Hamiltonian system is

$$y' = \sigma(y)J^{-1}\nabla H(y) + (H(y) - H_0)J^{-1}\nabla\sigma(y). \quad (2.4)$$

Compared to (2.1) we have introduced a perturbation, which vanishes along the solution of the Hamiltonian system passing through y_0 , but which makes the system Hamiltonian.

Time transformations such as in (2.3) are used in classical mechanics for an analytic treatment of Hamiltonian systems (Levi-Civita (1906, 1920), where (2.3) is called the “Darboux–Sundman transformation”, see Sundman (1912)). Zare & Szebehely (1975) consider such time transformations for numerical purposes (without taking care of symplecticity). Waldvogel & Spirig (1995) apply the transformations proposed by Levi-Civita to Hill’s lunar problem and solve the transformed equations by composition methods in order to preserve the symplectic structure. The following general procedure was proposed independently by Hairer (1997) and Reich (1999).

Algorithm 2.1. *Apply an arbitrary symplectic one-step method with constant step size ε to the Hamiltonian system (2.4), augmented by $t' = \sigma(y)$. This yields numerical approximations (y_n, t_n) with $y_n \approx y(t_n)$.*

Although this algorithm yields numerical approximations on a non-equidistant grid, it can be considered as a fixed step size, symplectic method applied to a different Hamiltonian system. This interpretation allows one to apply the standard techniques for the study of its long-time behaviour.

A disadvantage of this algorithm is that for separable Hamiltonians $H(p, q) = T(p) + U(q)$ the transformed Hamiltonian (2.3) is no longer separable. Hence, methods that are explicit for separable Hamiltonians are not explicit in the implementation of Algorithm 2.1. The following examples illustrate that this disadvantage can be partially overcome for the important case of Hamiltonian functions

$$H(p, q) = \frac{1}{2} p^T M^{-1} p + U(q), \quad (2.5)$$

where M is a constant symmetric matrix.

Example 2.2 (Symplectic Euler with p -Independent Step Size Function). For step size functions $\sigma(q)$ the symplectic Euler method, applied with constant step size ε to (2.4), reads

$$\begin{aligned} p_{n+1} &= p_n - \varepsilon \sigma(q_n) \nabla U(q_n) - \varepsilon \left(\frac{1}{2} p_{n+1}^T M^{-1} p_{n+1} + U(q_n) - H_0 \right) \nabla \sigma(q_n) \\ q_{n+1} &= q_n + \varepsilon \sigma(q_n) M^{-1} p_{n+1} \end{aligned}$$

and yields an approximation at $t_{n+1} = t_n + \varepsilon \sigma(q_n)$. The first equation is non-linear (quadratic) in p_{n+1} . Introducing the scalar quantity $\beta := \|p_{n+1}\|_M^2 := p_{n+1}^T M^{-1} p_{n+1}$, it reduces to the scalar quadratic equation

$$\beta = \left\| p_n - \varepsilon \sigma(q_n) \nabla U(q_n) - \varepsilon \left(\frac{\beta}{2} + U(q_n) - H_0 \right) \nabla \sigma(q_n) \right\|_M^2$$

which can be solved directly. The numerical solution (p_{n+1}, q_{n+1}) is then given explicitly.

Choices of Step Size Functions. Sometimes suitable functions $\sigma(p, q)$ are known a priori. For example, for the two-body problem one can take $\sigma(p, q) = \|q\|^\alpha$, e.g., $\alpha = 2$, or $\alpha = 3/2$ to preserve the scaling invariance (Budd & Piggott 2003), so that smaller step sizes are taken when the two bodies are close.

An interesting choice, which does not require any a priori knowledge of the solution, is $\sigma(y) = \|f(y)\|^{-1}$. The solution of (2.1) then satisfies $\|y'(\tau)\| = 1$ (arc-length parameterization) and we get approximations y_n that are nearly equidistant in the phase space. Such time transformations have been proposed by McLeod & Sanz-Serna (1982) for graphical reasons and by Huang & Leimkuhler (1997). For a Hamiltonian system with $H(p, q)$ given by (2.5), it is thus natural to consider

$$\sigma(p, q) = \left(\frac{1}{2} p^T M^{-1} p + \nabla U(q)^T M^{-1} \nabla U(q) \right)^{-1/2}. \quad (2.6)$$

We have chosen this particular norm, because it leaves the expression (2.6) invariant with respect to linear coordinate changes $q \mapsto Aq$ (implying $p \mapsto A^{-T}p$). Exploiting the fact that the Hamiltonian (2.5) is constant along solutions, the step size function (2.6) can be replaced by the p -independent function

$$\sigma(q) = \left((H_0 - U(q)) + \nabla U(q)^T M^{-1} \nabla U(q) \right)^{-1/2}. \quad (2.7)$$

The use of (2.6) and (2.7) gives nearly identical results, but (2.7) is easier to implement. If we are interested in an output that is approximatively equidistant in the q -space, we can take

$$\sigma(q) = (H_0 - U(q))^{-1/2}. \quad (2.8)$$

Example 2.3 (Störmer–Verlet Scheme with p -Independent Step Size Function).

For a step size function $\sigma(q)$ the Störmer–Verlet scheme gives

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{\varepsilon}{2} \sigma(q_n) \nabla U(q_n) - \frac{\varepsilon}{2} \left(H(p_{n+1/2}, q_n) - H_0 \right) \nabla \sigma(q_n) \\ q_{n+1} &= q_n + \frac{\varepsilon}{2} (\sigma(q_n) + \sigma(q_{n+1})) M^{-1} p_{n+1/2} \\ p_{n+1} &= p_{n+1/2} - \frac{\varepsilon}{2} \sigma(q_{n+1}) \nabla U(q_{n+1}) \\ &\quad - \frac{\varepsilon}{2} \left(H(p_{n+1/2}, q_{n+1}) - H_0 \right) \nabla \sigma(q_{n+1}). \end{aligned} \quad (2.9)$$

The first equation is essentially the same as that for the symplectic Euler method, and it can be solved for $p_{n+1/2}$ as explained in Example 2.2. The second equation is implicit in q_{n+1} , but it is sufficient to solve the scalar equation

$$\gamma = \sigma \left(q_n + \frac{\varepsilon}{2} (\sigma(q_n) + \gamma) M^{-1} p_{n+1/2} \right) \quad (2.10)$$

for $\gamma = \sigma(q_{n+1})$. Newton iterations can be efficiently applied, because $\nabla \sigma(q)$ is available already. The last equation (for p_{n+1}) is explicit. This variable step size Störmer–Verlet scheme gives approximations at t_n , where

$$t_{n+1} = t_n + \frac{\varepsilon}{2} (\sigma(q_n) + \sigma(q_{n+1})).$$

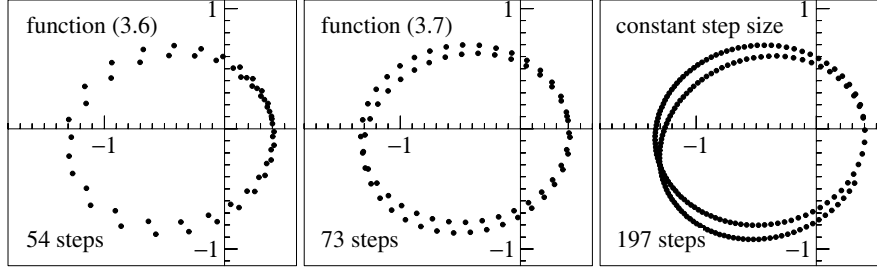


Fig. 2.1. Various step size strategies for the Störmer–Verlet scheme (Example 2.3) applied to the perturbed Kepler problem (1.2) on the interval $[0, 10]$ (approximately two periods)

In Fig. 2.1 we illustrate how the different step size functions influence the position of the output points. We apply the Störmer–Verlet method of Example 2.3 to the perturbed Kepler problem (1.2) with initial values, perturbation parameter, and eccentricity as in Sect. VIII.1. As step size functions we use (2.7), (2.8), and constant step size $\sigma(q) \equiv 1$. For all three choices of $\sigma(q)$ we have adjusted the parameter ε in such a way that the maximal error in the Hamiltonian is close to 0.01. The step size strategy (2.7) is apparently the most efficient one. For this strategy, we observe that the output points in the q -plane concentrate in regions where the velocity is large, while the constant step size implementation shows the opposite behaviour.

VIII.2.2 Reversible Integration

For ρ -reversible differential equations $\dot{y} = f(y)$, i.e., $f(\rho y) = -\rho f(y)$ for all y , the time transformed problem (2.1) remains ρ -reversible if

$$\sigma(\rho y) = \sigma(y). \quad (2.11)$$

This condition is not very restrictive and is satisfied by many important time transformations. In particular, (2.11) holds for the arc length parameterization $\sigma(y) = \|f(y)\|^{-1}$ if ρ is orthogonal. Consequently, it makes sense to apply symmetric, reversible numerical methods with constant step size ε directly to the system (2.1).

However, similar to the symplectic integration of Sect. VIII.2.1, there is a serious disadvantage. For separable differential equations (i.e., problems that can be split as $\dot{p} = f_1(q)$, $\dot{q} = f_2(p)$) and for non-constant $\sigma(p, q)$ the transformed system (2.1) is no longer separable. Hence, methods that are explicit for separable problems are not necessarily explicit for (2.1).

Example 2.4 (Adaptive Störmer–Verlet Method). We consider a Hamiltonian system with separable Hamiltonian (2.5), and we apply the Störmer–Verlet scheme to (2.1). This yields (Huang & Leimkuhler 1997)

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{\varepsilon}{2} s_n \nabla U(q_n) \\ q_{n+1} &= q_n + \frac{\varepsilon}{2} (s_n + s_{n+1}) M^{-1} p_{n+1/2} \\ p_{n+1} &= p_{n+1/2} - \frac{\varepsilon}{2} s_{n+1} \nabla U(q_{n+1}), \end{aligned} \quad (2.12)$$

where $s_n = \sigma(p_{n+1/2}, q_n)$ and $s_{n+1} = \sigma(p_{n+1/2}, q_{n+1})$ (notice that the s_{n+1} of the current step is not the same as the s_n of the subsequent step, if $\sigma(p, q)$ depends on p). The values (p_{n+1}, q_{n+1}) are approximations to the solution at t_n , where

$$t_{n+1} = t_n + \frac{\varepsilon}{2}(s_n + s_{n+1}).$$

For a p -independent step size function s , method (2.12) corresponds to that of Example 2.3, where the terms involving $\nabla\sigma(q)$ are removed. The implicitness of (2.12) is comparable to that of the method of Example 2.3. Completely explicit variants of this method will be discussed in the next section.

We conclude this section with a brief comparison of the variable step size Störmer–Verlet methods of Examples 2.3 and 2.4. Method (2.12) is easier to implement and more efficient when the step size function $\sigma(p, q)$ is expensive to evaluate. In a few numerical comparisons we observed, however, that the error in the Hamiltonian and in the solution is in general larger for method (2.12), and that the method (2.9) becomes competitive when $\sigma(p, q)$ is p -independent and easy to evaluate. A similar observation in favour of method (2.9) has been made by Calvo, López-Marcos & Sanz-Serna (1998).

VIII.3 Structure-Preserving Step Size Control

The disappointing long-time behaviour in Fig. 1.1 of the variable step size implementation of the Störmer–Verlet scheme is due to lack of reversibility. Indeed, for a ρ -reversible differential equation the step size $h_{n+1/2}$ taken for stepping from y_n to y_{n+1} should be the same as that when stepping from ρy_{n+1} to ρy_n (cf. Fig. V.1.1). The strategy of Sect. VIII.1, for which the step size depends on information of the preceding step, cannot guarantee such a property.

VIII.3.1 Proportional, Reversible Controllers

Following a suggestion of Stoffer (1988) we consider step sizes depending only on information of the present step, i.e., being *proportional* to some function of the actual state. This leads to the algorithm

$$y_{n+1} = \Phi_{h_{n+1/2}}(y_n), \quad h_{n+1/2} = \varepsilon s(y_n, \varepsilon), \quad (3.1)$$

where $\Phi_h(y)$ is a one-step method for $\dot{y} = f(y)$, and ε is a small parameter. For theoretical investigations it is useful to consider the mapping

$$\Psi_\varepsilon(y) := \Phi_{\varepsilon s(y, \varepsilon)}(y). \quad (3.2)$$

This is a one-step discretization, consistent with $y' = s(y, 0)f(y)$, and applied with constant step size ε . Consequently, all results concerning the long-time integration

with constant steps (e.g., backward error analysis of Chap. IX), and the definitions of symmetry and reversibility can be extended in a straightforward way.

Symmetry. We call the algorithm (3.1) symmetric, if $\Psi_\varepsilon(y)$ is symmetric, i.e., $\Psi_\varepsilon = \Psi_{-\varepsilon}^{-1}$. In the case of a symmetric Φ_h this is equivalent to

$$s(\hat{y}, -\varepsilon) = s(y, \varepsilon) \quad \text{with} \quad \hat{y} = \Phi_{\varepsilon s(y, \varepsilon)}(y). \quad (3.3)$$

Reversibility. The algorithm (3.1) is called ρ -reversible if, when applied to a ρ -reversible differential equation, $\Psi_\varepsilon(y)$ is ρ -reversible, i.e., $\rho \circ \Psi_\varepsilon = \Psi_\varepsilon^{-1} \circ \rho$ (cf. Definition V.1.2). If the method Φ_h is ρ -reversible then this is equivalent to

$$s(\rho^{-1}\hat{y}, \varepsilon) = s(y, \varepsilon) \quad \text{with} \quad \hat{y} = \Phi_{\varepsilon s(y, \varepsilon)}(y). \quad (3.4)$$

Example 3.1. Aiming at step sizes $h \approx \varepsilon \sigma(y)$ (cf. (2.2)), Hut, Makino & McMillan (1995) propose the use of $s(y, \varepsilon) = \frac{1}{2}(\sigma(y) + \sigma(\hat{y}))$ where, as in Sect. VIII.2, $\sigma(y)$ is some function that uses an a priori knowledge of the solution of the differential equation. Notice that, because of $\hat{y} = \Phi_{\varepsilon s(y, \varepsilon)}(y)$, the value of $s(y, \varepsilon)$ is defined by an implicit relation. Condition (3.3) is satisfied whenever $\Phi_h(y)$ is symmetric, and (3.4) is satisfied whenever $\Phi_h(y)$ is ρ -reversible and $\sigma(\rho y) = \sigma(y)$ holds. For a proof of these statements one shows that $s(\hat{y}, -\varepsilon)$ and $s(y, \varepsilon)$ (resp. $s(\rho^{-1}\hat{y}, \varepsilon)$ and $s(y, \varepsilon)$) are solution of the same nonlinear equation.

How can we find suitable step size functions $s(y, \varepsilon)$ which satisfy all these properties, and which do not require any a priori knowledge of the solution? In a remarkable publication, Stoffer (1995) gives the key to the answer of this question. He simply proposes to choose the step size h in such a way that the local error estimate satisfies $err = Tol$ (in contrast to $err \leq Tol$ for the standard strategy). Let us explain this idea in some more detail for Runge–Kutta methods.

Example 3.2 (Symmetric, Variable Step Size Runge–Kutta Methods). For the numerical solution of $\dot{y} = f(y)$ we consider Runge–Kutta methods

$$Y_i = y_n + h \sum_{j=1}^s a_{ij} f(Y_j), \quad y_{n+1} = y_n + h \sum_{i=1}^s b_i f(Y_i), \quad (3.5)$$

with coefficients satisfying $a_{s+1-i, s+1-j} + a_{ij} = b_j$ for all i, j . Such methods are symmetric and reversible (cf. Theorem V.2.3). A common approach for step size control is to consider an embedded method $\hat{y}_{n+1} = y_n + h \sum_{i=1}^s \hat{b}_i f(Y_i)$ (which has the same internal stages Y_i) and to take the difference $y_{n+1} - \hat{y}_{n+1}$, i.e.,

$$D(y_n, h) = h \sum_{i=1}^s e_i f(Y_i) \quad (3.6)$$

with $e_i = b_i - \hat{b}_i$, as indicator of the local error. For methods where $Y_i \approx y(t_n + c_i h)$ (e.g., collocation or discontinuous collocation) one usually computes the coefficients e_i from a nontrivial solution of the homogeneous linear system

$$\sum_{i=1}^s e_i c_i^{k-1} = 0 \quad \text{for } k = 1, \dots, s-1. \quad (3.7)$$

This yields $D(y_n, h) = \mathcal{O}(h^r)$ with r close to s . According to the suggestion of Stoffer (1995) we determine the step size $h_{n+1/2}$ such that

$$\|D(y_n, h_{n+1/2})\| = \text{tol}. \quad (3.8)$$

A Taylor expansion around $h = 0$ shows that $D(y, h) = d_r(y)h^r + \mathcal{O}(h^{r+1})$ with some $r \geq 1$. We assume $\|d_r(y)\| \neq 0$ and we put $\varepsilon = \text{tol}^{1/r}$, so that $h_{n+1/2}$ from (3.8) can be expressed by a smooth function $s(y, \varepsilon)$ as (3.1).

To satisfy the *symmetry* relation (3.3) we determine the e_i such that

$$e_{s+1-i} = e_i \quad \text{for all } i \quad \text{or} \quad e_{s+1-i} = -e_i \quad \text{for all } i \quad (3.9)$$

(Hairer & Stoffer 1997). If the Runge–Kutta method is symmetric, this then implies

$$\|D(y_n, h)\| = \|D(y_{n+1}, -h)\| \quad \text{with} \quad y_{n+1} = \Phi_h(y_n). \quad (3.10)$$

This follows from the fact that the internal stage vectors Y_i of the step from y_n to y_{n+1} and the stage vectors \bar{Y}_i of the step from y_{n+1} to y_n (negative step size $-h$) are related by $\bar{Y}_i = Y_{s+1-i}$. The step size determined by (3.8) is thus the same for both steps and, consequently, condition (3.3) holds.

The *reversibility* requirement (3.4) is a consequence of

$$\|D(y_n, h)\| = \|D(\rho^{-1}y_{n+1}, h)\| \quad \text{with} \quad y_{n+1} = \Phi_h(y_n) \quad (3.11)$$

which is satisfied for orthogonal mappings ρ (i.e., $\rho^T \rho = I$). This is seen as follows: applying Φ_h to $\rho^{-1}y_{n+1}$ gives $\rho^{-1}y_n$, and the internal stages are $\bar{Y}_i = \rho^{-1}Y_{s+1-i}$. Hence, we have from (3.9) that $D(\rho^{-1}y_{n+1}, h) = \pm \rho^{-1}D(y_n, h)$, and (3.11) follows from the orthogonality of ρ .

A simple special case is the trapezoidal rule

$$y_{n+1} = y_n + \frac{h_{n+1/2}}{2} (f(y_n) + f(y_{n+1})) \quad (3.12)$$

combined with

$$D(y_n, h) = \frac{h}{2} (f(y_{n+1}) - f(y_n)).$$

The scalar nonlinear equation (3.8) for $h_{n+1/2}$ can be solved in tandem with the nonlinear system (3.12).

Example 3.3 (Symmetric, Variable Step Size Störmer–Verlet Scheme). The strategy of Example 3.2 can be extended in a straightforward way to partitioned Runge–Kutta methods. For example, for the second order symmetric Störmer–Verlet scheme (I.1.17), applied to the problem $\dot{q} = p$, $\dot{p} = -\nabla U(q)$, we can take

$$D(p_n, q_n, h) = \frac{h}{2} \begin{pmatrix} \nabla U(q_{n+1}) - \nabla U(q_n) \\ h(\nabla U(q_{n+1}) + \nabla U(q_n)) \end{pmatrix}$$

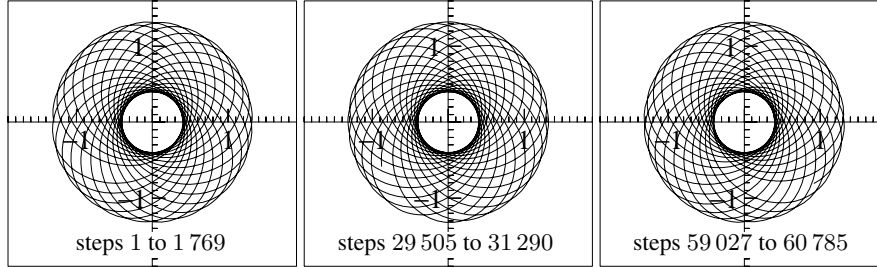


Fig. 3.1. Störmer–Verlet scheme applied with the symmetric adaptive step size strategy of Example 3.3 ($Tol = 0.01$); the three pictures have the same meaning as in Fig. 1.1

as error indicator. The first component is just the difference of the Störmer–Verlet solution and the numerical approximation obtained by the symplectic Euler method. The second component is a symmetrized version of it.

We apply this method with $h_{n+1/2}$ determined by (3.8) and $Tol = 0.01$ to the perturbed Kepler problem (1.2) with initial values as in Fig. 1.1. The result is given in Fig. 3.1. We identify a correct qualitative behaviour (compared to the wrong behaviour for the standard step size strategy in Fig. 1.1). It should be mentioned that the work for solving the scalar equation (3.8) for $h_{n+1/2}$ is not negligible, because the Störmer–Verlet scheme is explicit. Solving this equation iteratively, every iteration requires one force evaluation $\nabla U(q)$. An efficient solver for this scalar nonlinear equation should be used.

A Two-Step Proportional Controller. With the aim of obtaining a completely explicit integrator, Huang & Leimkuhler (1997) propose the use of two-term recurrence relations for the step size sequence, see also Holder, Leimkuhler & Reich (2001). Instead of using a relation between $h_{n+1/2}$, y_n and y_{n+1} (cf. Example 3.1) which is necessarily implicit, it is suggested to use a symmetric relation between $h_{n-1/2}$, $h_{n+1/2}$, and y_n , which then is explicit. In particular, with the notation $h_{n+1/2} = \varepsilon s_{n+1/2}$, it is proposed to use the two-term recurrence relation

$$\frac{1}{s_{n+1/2}} + \frac{1}{s_{n-1/2}} = \frac{2}{\sigma(y_n)}, \quad (3.13)$$

starting with $s_{1/2} = \sigma(y_0)$. In combination with the Störmer–Verlet method for separable Hamiltonians, this algorithm is completely explicit, and the authors report an excellent performance for realistic problems.

A rigorous analysis of the long-time behaviour of this variable step size Störmer–Verlet method is much more difficult. The results of Chapters IX and XI cannot be applied, because it is not a one-step mapping $y_n \mapsto y_{n+1}$. The analysis of Cirilli, Hairer & Leimkuhler (1999) shows that, similar to weakly stable multistep methods (Chap. XV), the numerical solution and the step size sequence contain oscillatory terms. Although these oscillations are usually very small (and hardly visible), it seems difficult to get rigorous estimates for them.

VIII.3.2 Integrating, Reversible Controllers

All variable step size approaches of this chapter are based on some time transformation $t \leftrightarrow \tau$ given by $\frac{dt}{d\tau} = \sigma(y)$ so that the differential equation, expressed in the new time variable τ , becomes

$$y' = \frac{1}{z} f(y), \quad z \sigma(y) = 1. \quad (3.14)$$

In Sect. VIII.2 we insert $z^{-1} = \sigma(y)$ into the differential equation and apply a numerical method to $y' = \sigma(y)f(y)$. In Sect. VIII.3.1 we first discretize the algebraic relation $z\sigma(y) = 1$ expressing $z_{n+1/2}$ in terms of y_n and y_{n+1} , and then apply a one-step method to the differential equation in (3.14) assuming $z = z_{n+1/2}$ being constant.

In the present section we first differentiate the algebraic relation of (3.14) with respect to τ . This yields by Leibniz' rule $z'\sigma(y) + z\nabla\sigma(y)^T y' = 0$ so that

$$z' = G(y) \quad \text{with} \quad G(y) = -\frac{1}{\sigma(y)} \nabla\sigma(y)^T f(y). \quad (3.15)$$

The idea of differentiating the constraint in (3.14) has been raised in Huang & Leimkuhler (1997), but soon abandoned in favour of the controller (3.13). The subsequent algorithm together with its theoretical justification is elaborated in Hairer & Söderlind (2004). The idea is to discretize first the differential equation in (3.15) and then to apply a one-step method to the problem (3.14) with constant z . The proposed algorithm is thus

$$\begin{aligned} z_{n+1/2} &= z_{n-1/2} + \varepsilon G(y_n) \\ y_{n+1} &= \Phi_{\varepsilon/z_{n+1/2}}(y_n) \end{aligned} \quad (3.16)$$

with $z_{1/2} = z_0 + \varepsilon G(y_0)/2$ and $z_0 = 1/\sigma(y_0)$. This algorithm is explicit whenever the underlying one-step method $\Phi_h(y)$ is explicit. It is called *integrating* controller, because the step size density is obtained by summing up small quantities.

For a theoretical analysis it is convenient to introduce $z_n = (z_{n+1/2} + z_{n-1/2})/2$ and to write (3.16) as a one-step method for the augmented system

$$y' = \frac{1}{z} f(y), \quad z' = G(y). \quad (3.17)$$

Notice that $I(y, z) = z \sigma(y)$ is a first integral of this system.

Algorithm 3.4. Let $\Phi_h(y)$ be a one-step method for $\dot{y} = f(y)$, $y(0) = y_0$. With $G(y)$ given by (3.15), $z_0 = 1/\sigma(y_0)$, and constant ε , we let

$$\begin{aligned} z_{n+1/2} &= z_n + \varepsilon G(y_n)/2 \\ y_{n+1} &= \Phi_{\varepsilon/z_{n+1/2}}(y_n) \\ z_{n+1} &= z_{n+1/2} + \varepsilon G(y_{n+1})/2. \end{aligned} \quad (3.18)$$

The values y_n approximate $y(t_n)$, where $t_{n+1} = t_n + \varepsilon/z_{n+1/2}$.

This algorithm has an interesting interpretation as Strang splitting for the solution of (3.17): it approximates the flow of $z' = G(y)$ with fixed y over a half-step $\varepsilon/2$; then applies the method Φ_ε to $y' = f(y)/z$ with fixed z ; finally, it computes a second half-step of $z' = G(y)$ with fixed y .

With the notation

$$\widehat{\Phi}_\varepsilon : \begin{pmatrix} y_n \\ z_n \end{pmatrix} \mapsto \begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} \quad \text{and} \quad \widehat{\rho} = \begin{pmatrix} \rho & 0 \\ 0 & 1 \end{pmatrix}. \quad (3.19)$$

the Algorithm 3.4 has the following properties:

- $\widehat{\Phi}_\varepsilon$ is symmetric whenever Φ_h is symmetric;
- $\widehat{\Phi}_\varepsilon$ is reversible with respect to $\widehat{\rho}$ whenever Φ_h is reversible with respect to ρ and $G(\rho y) = -G(y)$ (this is a consequence of $\sigma(\rho y) = \sigma(y)$).

These properties imply that standard techniques for constant step size implementations can be applied to $\widehat{\Phi}_\varepsilon$, and thus yield insight into the variable step size algorithm of this section. It will be shown in Chap. XI that when applied to integrable reversible systems there is no drift in the action variables and the global error grows only linearly with time. Moreover, the first integral $I(y, z) = z \sigma(y)$ of the system (3.17) is also well preserved (without drift) for such problems.

Example 3.5 (Variable Step Size Störmer–Verlet method). Consider a Hamiltonian system with separable Hamiltonian $H(p, q) = T(p) + U(q)$. Using the Störmer–Verlet method as basic method the above algorithm becomes (starting with $z_0 = 1/\sigma(y_0)$ and $z_{1/2} = z_0 + \varepsilon G(p_0, q_0)/2$)

$$\begin{aligned} z_{n+1/2} &= z_{n-1/2} + \varepsilon G(p_n, q_n) \\ p_{n+1/2} &= p_n - \varepsilon \nabla U(q_n)/(2z_{n+1/2}) \\ q_{n+1} &= q_n + \varepsilon \nabla T(p_{n+1/2})/z_{n+1/2} \\ p_{n+1} &= p_{n+1/2} - \varepsilon \nabla U(q_{n+1})/(2z_{n+1/2}). \end{aligned} \quad (3.20)$$

This method is explicit, symmetric and reversible as long as $G\rho = -G$, and computes approximations on a non-equidistant grid $\{t_n\}$ given by $t_{n+1} = t_n + \varepsilon/z_{n+1/2}$.

Let us apply this method to the perturbed Kepler problem with data and initial values as in the beginning of this chapter. Further, we select $\sigma(q) = (q^T q)^{\alpha/2}$ with $\alpha = 3/2$, so that the control function (3.15) becomes

$$G(p, q) = -\alpha p^T q / q^T q. \quad (3.21)$$

Figure 3.2 shows the error in the Hamiltonian along the numerical solution as well as the global error in the solution (fictive step size $\varepsilon = 0.02$). The error in the Hamiltonian is proportional to ε^2 without drift, and the global error grows linearly with time (in double logarithmic scale a linear growth corresponds to a line with slope one; such lines are drawn in grey). This is qualitatively the same behaviour as observed in constant step size implementations of symplectic methods.

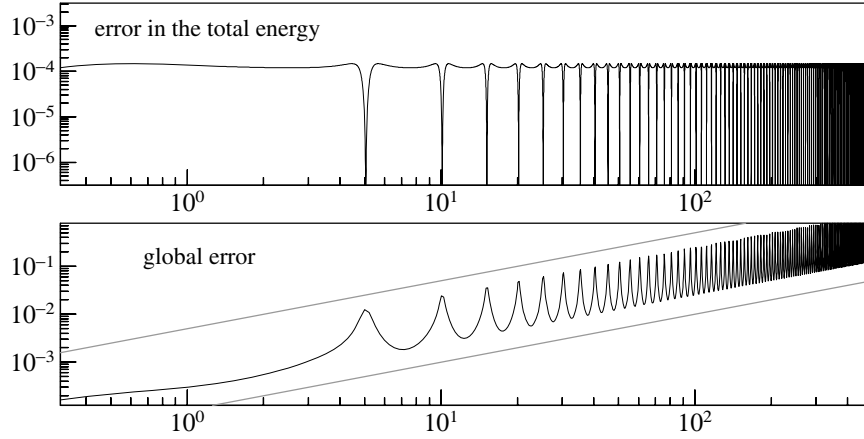


Fig. 3.2. Numerical Hamiltonian and global error as a function of time

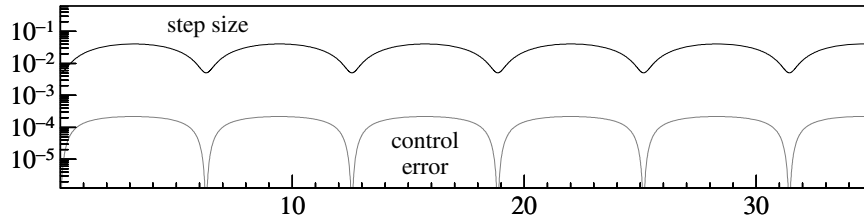


Fig. 3.3. Step sizes of the variable step size Störmer–Verlet method as a function of time, and the control error $z_n\sigma(q_n) - z_0\sigma(q_0)$ (grey curve)

Figure 3.3 shows the selected step sizes $h_{n+1/2} = \varepsilon/z_{n+1/2}$ as a function of time, and the control error $z_n\sigma(q_n) - z_0\sigma(q_0)$ in grey. Since its deviation from the constant value $z_0\sigma(q_0) = 1$ is small without any drift, the step density remains close to $1/\sigma(q)$. For an explanation of this excellent long-time behaviour we refer to Sect. XI.3.

VIII.4 Multiple Time Stepping

A completely different approach to variable step sizes will be described in this section. We are interested in situations where:

- many solution components of the differential equation vary slowly and only a few components have fast dynamics; or
- computationally expensive parts of the right-hand side do not contribute much to the dynamics of the solution.

In the first case it is tempting to use large step sizes for the slow components and small step sizes for the fast ones. Such integrators, called *multirate methods*, were

first formulated by Rice (1960) and Gear & Wells (1984). They were further developed by Günther & Rentrop (1993) in view of applications in electric circuit simulation, and by Engstler & Lubich (1997) with applications in astrophysics. Symmetric multirate methods are obtained from the approaches described below and are specially constructed by Leimkuhler & Reich (2001).

The second case suggests the use of methods that evaluate the expensive part of the vector field less often than the rest. This approach is called *multiple time stepping*. It was originally proposed for astronomy by Hayli (1967) and has become very popular in molecular dynamics simulations (Streett, Tildesley & Saville 1978, Grubmüller, Heller, Windemuth & Schulten 1991, Tuckerman, Berne & Martyna 1992). As noticed by Biesiadecki & Skeel (1993), one approach to such methods is within the framework of splitting and composition methods, which yields symmetric and symplectic methods. A second family of symmetric multiple time stepping methods results from the concept of using averaged force evaluations.

VIII.4.1 Fast-Slow Splitting: the Impulse Method

Consider a differential equation

$$\dot{y} = f(y), \quad f(y) = f^{[\text{slow}]}(y) + f^{[\text{fast}]}(y), \quad (4.1)$$

where the vector field is split into summands contributing to slow and fast dynamics, respectively, and where $f^{[\text{slow}]}(y)$ is more expensive to evaluate than $f^{[\text{fast}]}(y)$. Multirate methods can often be cast into this framework by collecting in $f^{[\text{slow}]}(y)$ those components of $f(y)$ which produce slow dynamics and in $f^{[\text{fast}]}(y)$ the remaining components.

Algorithm 4.1. For a given $N \geq 1$ and for the differential equation (4.1) a multiple time stepping method is obtained from

$$(\Phi_{h/2}^{[\text{slow}]})^* \circ (\Phi_{h/N}^{[\text{fast}]})^N \circ \Phi_{h/2}^{[\text{slow}]}, \quad (4.2)$$

where $\Phi_h^{[\text{slow}]}$ and $\Phi_h^{[\text{fast}]}$ are numerical integrators consistent with $\dot{y} = f^{[\text{slow}]}(y)$ and $\dot{y} = f^{[\text{fast}]}(y)$, respectively.

The method of Algorithm 4.1 is already stated in symmetrized form (Φ_h^* denotes the adjoint of Φ_h). It is often called the *impulse method*, because the slow part $f^{[\text{slow}]}$ of the vector field is used – impulse-like – only at the beginning and at the end of the step, whereas the many small substeps in between are concerned solely through integrating the fast system $\dot{y} = f^{[\text{fast}]}(y)$.

Lemma 4.2. Let $\Phi_h^{[\text{slow}]}$ be an arbitrary method of order 1, and $\Phi_h^{[\text{fast}]}$ a symmetric method of order 2. Then, the multiple time stepping algorithm (4.2) is symmetric and of order 2.

If $f^{[\text{slow}]}(y)$ and $f^{[\text{fast}]}(y)$ are Hamiltonian and if $\Phi_h^{[\text{slow}]}$ and $\Phi_h^{[\text{fast}]}$ are both symplectic, then the multiple time stepping method is also symplectic.

Proof. Due to the interpretation of multiple time stepping as composition methods the proof of these statements is obvious. \square

The order statement of Lemma 4.2 is valid for $h \rightarrow 0$, but should be taken with caution if the product of the step size h with a Lipschitz constant of the problem is not small (see Chap. XIII for a detailed analysis): it is *not* stated, and is not true in general for large N , that if h and h/N are the step sizes needed to integrate the slow and fast system, respectively, with an error bounded by ε , then the error of the combined scheme is $\mathcal{O}(\varepsilon)$.

The most important application of multiple time stepping is in Hamiltonian systems with a separable Hamiltonian

$$H(p, q) = T(p) + U(q), \quad U(q) = U^{[\text{slow}]}(q) + U^{[\text{fast}]}(q). \quad (4.3)$$

If we let the fast vector field correspond to $T(p) + U^{[\text{fast}]}(q)$ and the slow vector field to $U^{[\text{slow}]}(q)$, and if we apply the Störmer–Verlet method and exact integration, respectively, Algorithm 4.1 reads

$$\varphi_{h/2}^{[\text{slow}]} \circ \left(\varphi_{h/2N}^{[\text{fast}]} \circ \varphi_{h/N}^T \circ \varphi_{h/2N}^{[\text{fast}]} \right)^N \circ \varphi_{h/2}^{[\text{slow}]}, \quad (4.4)$$

where $\varphi_t^T, \varphi_t^{[\text{slow}]}, \varphi_t^{[\text{fast}]}$ are the exact flows corresponding to the Hamiltonian systems for $T(p), U^{[\text{slow}]}(q), U^{[\text{fast}]}(q)$, respectively. Notice that for $N = 1$ the method (4.4) reduces to the Störmer–Verlet scheme applied to the Hamiltonian system with $H(p, q)$. This is a consequence of the fact that $\varphi_t^{[\text{fast}]} \circ \varphi_t^{[\text{slow}]} = \varphi_t^U$ is the exact flow of the Hamiltonian system corresponding to $U(q)$ of (4.3). In the molecular dynamics literature, the method (4.4) is known as the Verlet-I method (Grubmüller et al. 1991, who consider the method with little enthusiasm) or r-RESPA method (Tuckerman et al. 1992, with much more enthusiasm).

Example 4.3. In order to illustrate the effect of multiple time stepping we choose a ‘solar system’ with two planets, i.e., with a Hamiltonian

$$H(p, q) = \frac{1}{2} \left(\frac{p_0^T p_0}{m_0} + \frac{p_1^T p_1}{m_1} + \frac{p_2^T p_2}{m_2} \right) - \frac{m_0 m_1}{\|q_0 - q_1\|} - \frac{m_0 m_2}{\|q_0 - q_2\|} - \frac{m_1 m_2}{\|q_1 - q_2\|},$$

where $m_0 = 1, m_1 = m_2 = 10^{-2}$ and initial values $q_0 = (0, 0), \dot{q}_0 = (0, 0), q_1 = (1, 0), \dot{q}_1 = (0, 1), q_2 = (4, 0), \dot{q}_2 = (0, 0.5)$. With these data, the motion of the two planets is nearly circular with periods close to 2π and 14π , respectively.

We split the potential as

$$U^{[\text{fast}]}(q) = -\frac{m_0 m_1}{\|q_0 - q_1\|}, \quad U^{[\text{slow}]}(q) = -\frac{m_0 m_2}{\|q_0 - q_2\|} - \frac{m_1 m_2}{\|q_1 - q_2\|},$$

and we apply the algorithm of (4.4) with $N = 1$ (Störmer–Verlet), $N = 4$, and $N = 8$. Since the evaluation of $\varphi_t^{[\text{slow}]}$ is about twice as expensive as $\varphi_t^{[\text{fast}]}$ and that of φ_t^T is of negligible cost, the computational work of applying (4.4) on a fixed interval is proportional to

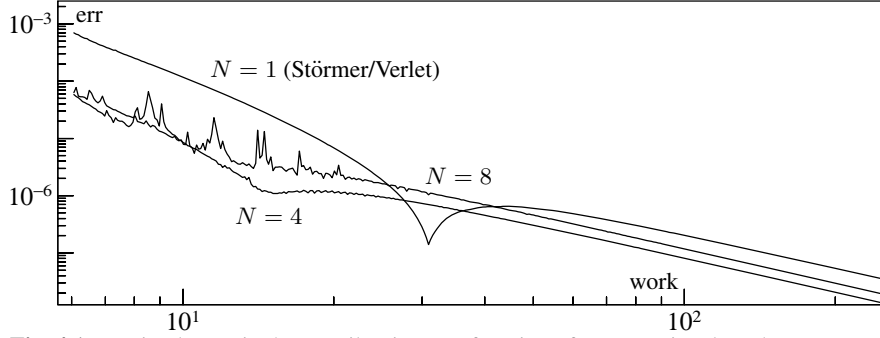


Fig. 4.1. Maximal error in the Hamiltonian as a function of computational work

$$\frac{2\pi}{h} \cdot \frac{(2+N)}{3}. \quad (4.5)$$

Our computations have shown that this measure of work corresponds very well to the actual cpu time.

We have solved this problem with many different step sizes h . Figure 4.1 shows the maximal error in the Hamiltonian (over the interval $[0, 200\pi]$) as a function of the computational work (4.5). We notice that the value $N = 4$ yields excellent results for relatively large as well as small step sizes. It noticeably improves the performance of the Störmer–Verlet method. If N becomes too large, an irregular behaviour for large step sizes is observed. Such “artificial resonances” are notorious for this method and have been discussed by Biesiadecki & Skeel (1993) for a similar experiment; also see Chap. XIII. For large N we also note a loss of accuracy for small step sizes. The optimal choice of N (which here is close to 4) depends on the problem and on the splitting into fast and slow parts, and has to be determined by experiment.

The multiple time stepping technique can be iteratively extended to problems with more than two different time scales. The idea is to split the ‘fast’ vector field of (4.1) into $f^{[\text{fast}]}(y) = f^{[ff]}(y) + f^{[fs]}(y)$, and to replace the method $\Phi_h^{[\text{fast}]}$ in Algorithm 4.1 with a multiple time stepping method. Depending on the problem, a significant gain in computer time may be achieved in this way.

Many more multiple time stepping methods that extend the above Verlet-I/r-RESPA/impulse method, have been proposed in the literature, most notably the mollified impulse method of García-Archilla, Sanz-Serna & Skeel (1999); see Sect. XIII.1.

VIII.4.2 Averaged Forces

A different approach to multiple time stepping arises from the idea of advancing the step with *averaged force evaluations*. We describe such a method for the second-order equation

$$\ddot{y} = f(y), \quad f(y) = f^{[\text{slow}]}(y) + f^{[\text{fast}]}(y). \quad (4.6)$$

The exact solution satisfies

$$y(t+h) - 2y(t) + y(t-h) = h^2 \int_{-1}^1 (1-|\theta|) f(y(t+\theta h)) d\theta,$$

where the integral on the right-hand side represents a weighted average of the force along the solution, which is now going to be approximated. At $t = t_n$, we replace

$$f(y(t_n + \theta h)) \approx f^{[\text{slow}]}(y_n) + f^{[\text{fast}]}(u(\theta h))$$

where $u(\tau)$ is a solution of the differential equation

$$\ddot{u} = f^{[\text{slow}]}(y_n) + f^{[\text{fast}]}(u). \quad (4.7)$$

We then have

$$h^2 \int_{-1}^1 (1-|\theta|) \left(f^{[\text{slow}]}(y_n) + f^{[\text{fast}]}(u(\theta h)) \right) d\theta = u(h) - 2u(0) + u(-h).$$

The velocities are treated similarly, starting from the identity

$$\dot{y}(t+h) - \dot{y}(t-h) = h \int_{-1}^1 f(y(t+\theta h)) d\theta.$$

A Symmetric Two-Step Method. For the differential equation (4.7) we assume the initial values

$$u(0) = y_n, \quad \dot{u}(0) = \dot{y}_n. \quad (4.8)$$

This initial value problem is solved numerically, e.g., by the Störmer–Verlet method with a smaller step size $\pm h/N$ on the interval $[-h, h]$, yielding numerical approximations $u_N(\pm h)$ and $v_N(\pm h)$ to $u(\pm h)$ and $\dot{u}(\pm h)$, respectively. Note that no further evaluations of $f^{[\text{slow}]}$ are needed for the computation of $u_N(\pm h)$ and $v_N(\pm h)$. This finally gives the symmetric two-step method (Hochbruck & Lubich 1999a)

$$\begin{aligned} y_{n+1} - 2y_n + y_{n-1} &= u_N(h) - 2u_N(0) + u_N(-h) \\ \dot{y}_{n+1} - \dot{y}_{n-1} &= v_N(h) - v_N(-h). \end{aligned} \quad (4.9)$$

The starting values y_1 and \dot{y}_1 are chosen as $u_N(h)$ and $v_N(h)$ which correspond to (4.7) and (4.8) for $n = 0$.

A Symmetric One-step Method. An explicit one-step method with similar averaged forces is obtained when the initial values for (4.7) are chosen as

$$u(0) = y_n, \quad \dot{u}(0) = 0. \quad (4.10)$$

It may appear crude to take zero initial values for the velocity, but we remark that for linear $f^{[\text{fast}]}$ the averaged force $(u(h) - 2u(0) + u(-h))/h^2$ does not depend on

the choice of $\dot{u}(0)$. Moreover the solution then satisfies $u(-t) = u(t)$, so that the computational cost is halved. We again denote by $u_N(h) = u_N(-h)$ the numerical approximation to $u(h)$ obtained with step size $\pm h/N$ from a one-step method (e.g., from the Störmer–Verlet scheme). Because of (4.10) the averaged forces

$$F_n = \frac{1}{h^2} (u_N(h) - 2u_N(0) + u_N(-h)) = \frac{2}{h^2} (u_N(h) - u_N(0))$$

now depend only on y_n and not on the velocity \dot{y}_n . In trustworthy Verlet manner, the scheme $y_{n+1} - 2y_n + y_{n-1} = h^2 F_n$ can be written as the one-step method

$$\begin{aligned} v_{n+1/2} &= v_n + \frac{h}{2} F_n \\ y_{n+1} &= y_n + h v_{n+1/2} \\ v_{n+1} &= v_{n+1/2} + \frac{h}{2} F_{n+1}. \end{aligned} \tag{4.11}$$

The auxiliary variables v_n can be interpreted as averaged velocities: we have

$$v_n = \frac{y_{n+1} - y_{n-1}}{2h} \approx \frac{y(t_{n+1}) - y(t_{n-1}))}{2h} = \frac{1}{2} \int_{-1}^1 \dot{y}(t_n + \theta h) d\theta.$$

This average may differ substantially from $\dot{y}(t_n)$ if the solution is highly oscillatory in $[-h, h]$. In the experiments of this section it turned out that the choice $v_0 = \dot{y}_0$ and $\dot{y}_n = v_n$ as velocity approximations gives excellent results.

In a multirate context, symmetric one-step schemes using averaged forces were studied by Hochbruck & Lubich (1999b), Nettesheim & Reich (1999), and Leimkuhler & Reich (2001). A closely related approach for problems with multiple time scales is the heterogeneous multiscale method by E (2003) and Engquist & Tsai (2005).

Example 4.4. We add a satellite of mass $m_3 = 10^{-4}$ to the three body-problem of Example 4.3. It moves rapidly around the planet number one. The initial positions and velocities are $q_3 = (1.01, 0)$ and $p_3 = (0, 0)$. We split the potential as

$$U^{[\text{fast}]}(q) = -\frac{m_1 m_3}{\|q_1 - q_3\|}, \quad U^{[\text{slow}]}(q) = -\sum_{\substack{i < j \\ (i,j) \neq (1,3)}} \frac{m_i m_j}{\|q_i - q_j\|},$$

and we apply the methods (4.9), (4.11), and the impulse method (4.4). Since the sum in $U^{[\text{slow}]}$ contains 5 terms, the computational work is proportional to

$$\begin{aligned} \frac{5 + N}{6h} & \quad \text{for methods (4.11) and (4.4)} \\ \frac{6 + 2N}{6h} & \quad \text{for method (4.9).} \end{aligned}$$

For each of the methods we have optimized the number N of small steps. We obtained a flat minimum near $N = 40$ for (4.9) and (4.4), and a more pronounced minimum at $N = 12$ for (4.11). Figure 4.2 shows the errors at $t = 10$ in the positions and in the Hamiltonian as a function of the computational work.

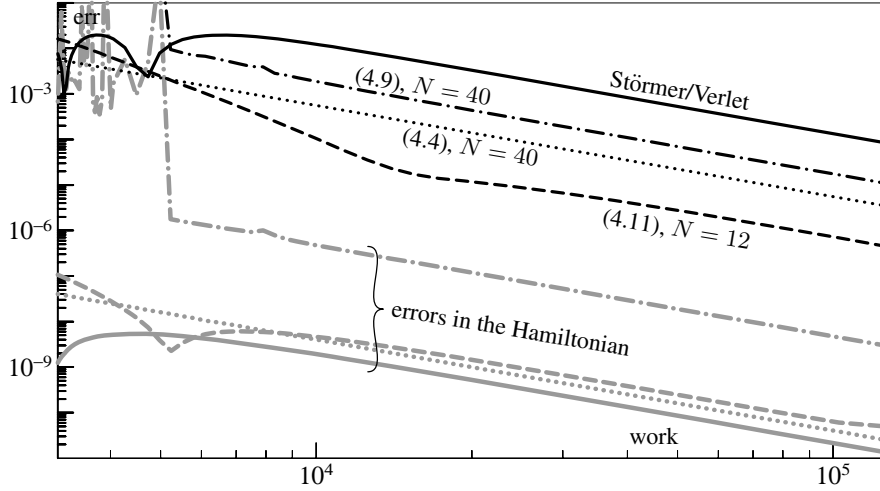


Fig. 4.2. Errors in position and in the Hamiltonian as a function of the computational work; the classical Störmer–Verlet method, the impulse method (4.4), and the averaged force methods (4.11) and (4.9). The errors in the Hamiltonian are indicated by grey lines (same linestyle)

The error in the position is largest for the Störmer–Verlet method and significantly smallest for the one-step averaged-force method (4.11). The errors in the velocities are about a factor 100 larger for all methods. They are not included in the figure. The error in the Hamiltonian is very similar for all methods with the exception of the two-step averaged-force method (4.9), for which it is much larger.

VIII.5 Reducing Rounding Errors

... the idea is to capture the rounding errors and feed them back into the summation. (N.J. Higham 1993)

All numerical methods for solving ordinary differential equations require the computation of a recursion of the form

$$y_{n+1} = y_n + \delta_n, \quad (5.1)$$

where δ_n , the increment, is usually smaller in magnitude than the approximation y_n to the solution. In this situation the rounding errors caused by the computation of δ_n are in general smaller than those due to the addition in (5.1).

A first attempt at reducing the accumulation of rounding errors (in fixed-point arithmetic for his Runge–Kutta code) was due to Gill (1951). Kahan (1965) and Möller (1965) both extended this idea to floating point arithmetic. The resulting algorithm is nowadays called ‘compensated summation’, and a particularly nice presentation and analysis is given by N. Higham (1993). In the following algorithm we assume that y_n is a scalar; vector valued recursions are treated componentwise.

Algorithm 5.1 (Compensated Summation). *Let y_0 and $\{\delta_n\}_{n \geq 0}$ be given and put $e = 0$. Compute y_1, y_2, \dots from (5.1) as follows:*

```

for  $n = 0, 1, 2, \dots$  do
     $a = y_n$ 
     $e = e + \delta_n$ 
     $y_{n+1} = a + e$ 
     $e = e + (a - y_{n+1})$ 
end do

```

This algorithm can best be understood with the help of Fig. 5.1 (following the presentation of N. Higham (1993)). We present the mantissas of floating point numbers by boxes, for which the horizontal position indicates the exponent (for a large exponent the box is more to the left). The mantissas of y_n and e together represent the accurate value of y_n (notice that in the beginning $e = 0$). The operations of Algorithm 5.1 yield y_{n+1} and a new e , which together represent $y_{n+1} = y_n + \delta_n$. No digit of δ_n is lost in this way. With a standard summation the last digits of δ_n (those indicated by δ'' in Fig. 5.1) would have been missed.

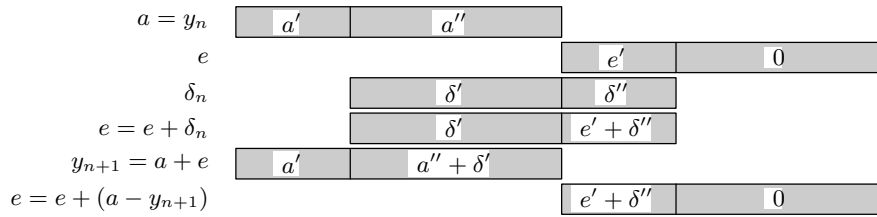


Fig. 5.1. Illustration of the technique of “compensated summation”

Numerical Experiment. We study the effect of compensated summation on the Kepler problem (I.2.2) (written as a first order system) with eccentricity $e = 0.6$ and initial values as in (I.2.11), so that the period of the elliptic orbit is exactly 2π . As the numerical integrator we take the composition method (V.3.13) of order 8 with the Störmer–Verlet scheme as basic integrator. We compute the numerical solution with step size $h = 2\pi/500$ once with standard update of the increment, once with compensated summation (both in double precision) and, in order to get a reference solution, we also perform the whole computation in quadruple precision. The difference between the double and quadruple precision computations gives us the rounding errors. Their Euclidean norms as a function of time are displayed in Fig. 5.2.

We see that throughout the whole integration interval the rounding errors of the standard implementation are nearly a factor of 100 larger than those of the implementation with compensated summation. This corresponds to the inverse of the step size or, more precisely, to the mean quotient between y_n and δ_n in (5.1). In Fig. 5.2 we have also included the pure global error of the method (without rounding errors) at integral multiples of the period 2π (hence no oscillations are visible). This is

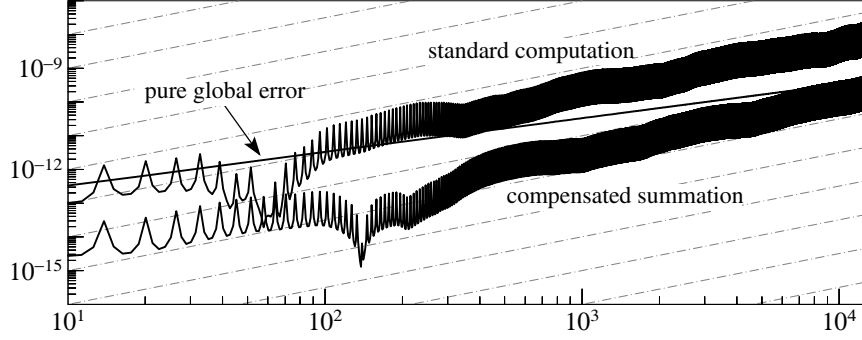


Fig. 5.2. Rounding errors and pure global error as a function of time; the parallel grey lines indicate a growth of $\mathcal{O}(t^{3/2})$

obtained as the difference of the numerical solution computed with quadruple precision and the exact solution. We observe a linear growth of the pure global error (this will be explained in Sect. X.3) and a growth like $\mathcal{O}(t^{3/2})$ due to the rounding errors. Thus, eventually the rounding errors will surpass the truncation errors, but this happens for the compensated summation only after some 1000 periods.

Probabilistic Explanation of the Error Growth. Our aim is to explain the growth rate of rounding errors observed in Fig. 5.2. Denote by ε_k the vector of rounding errors produced during the computations in the k th step. Since the derivative of the flow $\varphi_t(y)$ describes the propagation of these errors, the accumulated rounding error at time $t = t_N$ ($t_k = kh$) is

$$\eta_t = \sum_{k=1}^N \varphi'_{t-t_k}(y_k) \varepsilon_k. \quad (5.2)$$

For the Kepler problem and, in fact, for all completely integrable differential equations (cf. Sect. X.1) the flow and its derivative grow at most linearly with time, i.e.,

$$\|\varphi'_{t-t_k}(y)\| \leq a + b(t - t_k) \quad \text{for } t \geq t_k. \quad (5.3)$$

Using $\varepsilon_k = \mathcal{O}(\text{eps})$, where eps denotes the roundoff unit of the computer, an application of the triangle inequality to (5.2) yields $\eta_t = \mathcal{O}(t^2 \text{eps})$. From our experiment of Fig. 5.2 we see that such an estimate is too pessimistic.

For a better understanding of accumulated rounding errors over long time intervals we make use of probability theory. Such an approach has been developed in the classical book of Henrici (1962). We assume that the components ε_{ki} of ε_k are *random variables* with mean and variance

$$E(\varepsilon_{ki}) = 0, \quad \text{Var}(\varepsilon_{ki}) = C_{ki} \cdot \text{eps}^2,$$

and uniformly bounded $C_{ki} \leq C$. For simplicity we assume that all ε_{ki} are independent random variables. Replacing the matrix $\varphi_{t-t_k}(y_k)$ in (5.2) with $\varphi_{t-t_k}(y(t_k))$

and denoting its entries by w_{ijk} , the i th component of the accumulated rounding error (5.2) becomes

$$\eta_{ti} = \sum_{k=1}^N \sum_{j=1}^n w_{ijk} \varepsilon_{kj},$$

a linear combination of the random variables ε_{kj} . Elementary probability theory thus implies that

$$E(\eta_{ti}) = 0 \quad \text{and} \quad \text{Var}(\eta_{ti}) = \sum_{k=1}^N \sum_{j=1}^n w_{ijk}^2 \text{Var}(\varepsilon_{kj}).$$

Inserting the estimate (5.3) for w_{ijk} we get

$$\text{Var}(\eta_{ti}) \leq \sum_{k=1}^N (a + b(t - t_k))^2 \max_{j=1, \dots, n} \text{Var}(\varepsilon_{kj}) = \mathcal{O}\left(\frac{C}{h} t^3 \text{eps}^2\right).$$

Consequently, the Euclidean norm of the expected rounding error η_t is

$$\left(\sum_{i=1}^n \text{Var}(\eta_{ti})\right)^{1/2} = \mathcal{O}\left(\sqrt{\frac{C}{h}} t^{3/2} \text{eps}\right).$$

This is in excellent agreement with the results displayed in Fig. 5.2.

VIII.6 Implementation of Implicit Methods

Symplectic methods for general Hamiltonian equations are implicit, and so are symmetric methods for general reversible systems. Also, when we consider variable step size extensions as described in Sections VIII.3 and VIII.2, we are led to nonlinear equations. The efficient numerical solution of such nonlinear equations is the main difficulty in an implementation of implicit methods. Notice that in the context of geometric integration there is no need of ad-hoc strategies for step size and order selection, so that the remaining parts of a computer code are more or less straightforward.

In the following we discuss the numerical solution of the nonlinear system defined by an implicit Runge–Kutta method. We have the Gauss methods of Sect. II.1.3 in mind which are symplectic and symmetric. An extension of the ideas to partitioned Runge–Kutta methods and to Nyström methods is obvious. For simplicity of notation we consider autonomous differential equations $\dot{y} = f(y)$, and we write the nonlinear system of Definition II.1.1 in the form

$$Z_{in} - h \sum_{j=1}^s a_{ij} f(y_n + Z_{jn}) = 0, \quad i = 1, \dots, s. \quad (6.1)$$

The unknown variables are Z_{1n}, \dots, Z_{sn} , and the equivalence of the two formulations is via the relation $k_i = f(y_n + Z_{in})$. The numerical solution after one step can be expressed as

$$y_{n+1} = y_n + h \sum_{i=1}^s b_i f(y_n + Z_{in}). \quad (6.2)$$

For implicit Runge–Kutta methods the equations (6.1) represent a nonlinear system that has to be solved iteratively. We discuss the choice of good starting approximations for Z_{in} as well as different nonlinear equation solvers (fixed-point iteration, modified Newton methods).

VIII.6.1 Starting Approximations

The most simple approximations to the solution Z_{in} of (6.1) are $Z_{in}^0 = 0$ or $Z_{in}^0 = hc_i f(y_n)$ where $c_i = \sum_{j=1}^s a_{ij}$. They are, however, not very accurate and we will try to exploit the information of previous steps for improving them. There are essentially two possibilities: either use only the information of the last step $y_{n-1} \mapsto y_n$ (methods (A) and (B) below), or consider a fixed i and use the interpolation polynomial that passes through $Z_{i,n-l}$ for $l = 1, 2, \dots$ (method (C)). Let us separately discuss these two approaches.

(A) Use of Continuous Output. Consider the polynomial $w_{n-1}(t)$ of degree s that interpolates the values (t_{n-1}, y_{n-1}) and $(t_{n-1} + c_i h, Y_{i,n-1})$ for $i = 1, \dots, s$, where $Y_{i,n-1} = y_{n-1} + Z_{i,n-1}$ is the argument in (6.1) of the previous step. For collocation methods (such as Gauss methods) $w_{n-1}(t)$ is the collocation polynomial, and we know from Lemma II.1.6 that on compact intervals

$$w_{n-1}(t) - y(t) = \mathcal{O}(h^{q+1}) \quad (6.3)$$

with $q = s$, where $y(t)$ denotes the solution of $\dot{y} = f(y)$ satisfying $y(t_{n-1}) = y_{n-1}$. For Runge–Kutta methods that are not collocation methods, (6.3) holds with q defined by the condition $C(q)$ of (II.1.11). Since the solution of $\dot{y} = f(y)$ passing through $y(t_n) = y_n$ is $\mathcal{O}(h^{p+1})$ close to $y(t)$ with $p \geq q$, we have $w_n(t) = w_{n-1}(t) + \mathcal{O}(h^{q+1})$ and the computable value

$$Z_{in}^0 = Y_{in}^0 - y_n, \quad Y_{in}^0 = w_{n-1}(t_n + c_i h) \quad (6.4)$$

serves as starting approximation for (6.1) with an error of size $\mathcal{O}(h^{q+1})$. This approach is standard in variable step size implementations of implicit Runge–Kutta methods (cf. Sect. IV.8 of Hairer & Wanner (1996)). Since $w_{n-1}(t) - y_{n-1}$ is a linear combination of the $Z_{i,n-1} = Y_{i,n-1} - y_{n-1}$, it follows from (6.1) that it is also a linear combination of $hf(Y_{i,n-1})$, so that

$$Y_{in}^0 = y_{n-1} + h \sum_{j=1}^s \beta_{ij} f(Y_{j,n-1}). \quad (6.5)$$

For a constant step size implementation, the β_{ij} depend only on the method coefficients and can be computed in advance as the solution of the linear Vandermonde type system

$$\sum_{j=1}^s \beta_{ij} c_j^{k-1} = \frac{(1 + c_i)^k}{k}, \quad k = 1, \dots, s \quad (6.6)$$

(see Exercise 2). For collocation methods and for methods with $q \geq s - 1$ the coefficients β_{ij} from (6.6) are optimal in the sense that they are the only ones making (6.5) an s th order approximation to the solution of (6.1). For $q < s - 1$, more complicated order conditions have to be considered (Sand 1992).

(B) Starting Algorithms Using Additional Function Evaluations. In particular for high order methods where s is relatively large, a much more accurate starting approximation can be constructed with the aid of a few additional function evaluations. Such starting algorithms have been investigated by Laburta (1997), who presents coefficients for the Gauss methods up to order 8 in Laburta (1998).

The idea is to use starting approximations of the form

$$Y_{in}^0 = y_{n-1} + h \sum_{j=1}^s \beta_{ij} f(Y_{j,n-1}) + h \sum_{j=1}^m \nu_{ij} f(Y_{s+j,n-1}), \quad (6.7)$$

where $Y_{1,n-1}, \dots, Y_{s,n-1}$ are the internal stages of the basic implicit Runge–Kutta method (with coefficients c_i, a_{ij}, b_j), and the additional internal stages are computed from

$$Y_{s+i,n-1} = y_{n-1} + h \sum_{j=1}^{s+i-1} \mu_{ij} f(Y_{j,n-1}).$$

For a fixed i , we interpret Y_{in}^0 as the result of the explicit Runge–Kutta method with coefficients of the right tableau of

exact i th stage	approximate
$\begin{array}{c cc} c & A & \\ \mathbb{1} + c & B & A \\ \hline & b^T & a_i^T \end{array}$	$\begin{array}{c cc} c & A & \\ \mu & M_1 & M_2 \\ \hline & \beta_i^T & \nu_i^T \end{array}$

(6.8)

Here, $(M_1, M_2) = M = (\mu_{jk})$, $\mu_j = \sum_{k=1}^{s+j-1} \mu_{jk}$, and c, μ, β_i, ν_i are the vectors composed of $c_j, \mu_j, \beta_{ij}, \nu_{ij}$, respectively. The exact stage values Y_{in} are interpreted as the result of the Runge–Kutta method with coefficients given in the left tableau of (6.8). The entries of the vectors $\mathbb{1}, b$ and a_i are 1, b_j and a_{ij} , respectively, and B is the matrix whose rows are all equal to b^T .

If the order conditions (see Sect. III.1) for the two Runge–Kutta methods of (6.8) give the same result for all trees with $\leq r$ vertices, we get an approximation of order r , i.e., $Y_{in}^0 - Y_{in} = \mathcal{O}(h^{r+1})$. For the bushy tree $\tau_k = [\bullet, \dots, \bullet]$ with k vertices we have

$$\sum_{j=1}^s \beta_{ij} c_j^{k-1} + \sum_{j=1}^m \nu_{ij} \mu_j^{k-1} = \sum_{j=1}^s b_j c_j^{k-1} + \sum_{j=1}^s a_{ij} (1 + c_j)^{k-1}. \quad (6.9)$$

Notice that for collocation methods (such as the Gauss methods) the condition $C(s)$ reduces the right-hand expression of this equation to $(1 + c_i)^k/k$ for $k \leq s$. For $m = 0$, these conditions are thus equivalent to (6.6).

For the tree $[\tau_k] = [[\bullet, \dots, \bullet]]$ with $k + 1$ vertices we get the condition

$$\begin{aligned} \sum_{j,l=1}^s \beta_{ij} a_{jl} c_l^{k-1} + \sum_{j=1}^m \nu_{ij} \left(\sum_{l=1}^s \mu_{jl} c_l^{k-1} + \sum_{l=1}^m \mu_{j,s+l} \mu_l^{k-1} \right) \\ = \sum_{j,l=1}^s b_j a_{jl} c_l^{k-1} + \sum_{j,l=1}^s a_{ij} \left(b_l c_l^{k-1} + a_{jl} (1 + c_l)^{k-1} \right). \end{aligned} \quad (6.10)$$

We now assume that the Runge–Kutta method corresponding to the right tableau of (6.8) satisfies condition $C(s)$. This means that the method (c, A, b) is a collocation method, and that the coefficients μ_{ij} have to be computed from the linear system

$$\sum_{j=1}^{s+i-1} \mu_{ij} c_j^{k-1} = \frac{\mu_i^k}{k}, \quad k = 1, \dots, s. \quad (6.11)$$

The method corresponding to the left tableau of (6.8) then also satisfies $C(s)$. Consequently, the order conditions are simplified considerably, and it follows from Sect. III.1 that Y_{in}^0 is an approximation to the exact stage value Y_{in} of order $s + 1$ or $s + 2$ if the following conditions hold:

$$\begin{aligned} \text{order } s + 1 & \quad \text{if (6.9) for } k = 1, \dots, s + 1; \\ \text{order } s + 2 & \quad \text{if (6.9) for } k = 1, \dots, s + 2, \text{ and (6.10) for } k = s + 1. \end{aligned} \quad (6.12)$$

For an approximation of order $s + 1$ we put $m = 1$, we arbitrarily choose μ_1 , we compute μ_{1j} from (6.11), and the coefficients β_{ij} and ν_{i1} from (6.9) with $k = 1, \dots, s + 1$. A reasonable choice for the free parameter is $\mu_1 \in [1, 2]$ (in our computations we take $\mu_1 = 1.75$ for $s = 2, 4$, and $\mu_1 = 1.8$ for $s = 6$).¹

For an approximation of order $s + 2$ we put $m = 3$. One of the three additional function evaluations can be saved if we put $\mu_1 = 0$ and $\mu_2 = 1$. This implies $Y_{s+1,n-1} = y_{n-1}$ and $Y_{s+2,n-1} = y_n$, so that the evaluation of $f(Y_{s+1,n-1})$ is already available from computations for the preceding step (FSAL technique, “first same as last”). In our experiments we take $\mu_3 = 1.6$ for $s = 2$, $\mu_3 = 1.65$ for $s = 4$, and $\mu_3 = 1.75$ for $s = 6$. The coefficients $\mu_{ij}, \beta_{ij}, \nu_{ij}$ are then obtained as the solution of Vandermonde like linear systems.

For an implementation it is more convenient to work with the quantities $Z_{in}^0 = Y_{in}^0 - y_n$ and to write (6.7) in the form

¹ Laburta (1997) proposes to consider $m = 2$, $\mu_1 = 0$, $\mu_2 = 1$ (apart from the first step this also needs only one additional function evaluation per step), and to optimize free parameters by satisfying the order conditions for some trees with one order higher.

$$Z_{in}^0 = h \sum_{j=1}^s \alpha_{ij} f(Y_{j,n-1}) + h \sum_{j=1}^m \nu_{ij} f(Y_{s+j,n-1}) \quad (6.13)$$

with $\alpha_{ij} = \beta_{ij} - b_j$.

(C) Equistage Approximation. From the implicit function theorem, applied to the nonlinear system (6.1), we know that $Z_{in} = z(y_n, h)$, where the function $z(y, h)$ is as smooth as $f(y)$. Furthermore, since on compact intervals the global error of a one-step method permits an asymptotic expansion in powers of h , we have $y_{n-l} = y_N(t_{n-l}, h) + \mathcal{O}(h^{N+1})$ with $y_N(t, h) = y(t) + h^p e_p(t) + \dots + h^N e_N(t)$ (the value of N can be chosen arbitrarily large if $f(y)$ is sufficiently smooth). Consequently, $Z_{i,n-l}$ is $\mathcal{O}(h^{N+1})$ close to the smooth function $z(y_N(t, h), h)$ at $t = t_n - lh$. Let $\zeta_i(t)$ be the polynomial of degree $k-1$ defined by $\zeta_i(t_{n-l}) = Z_{i,n-l}$ for $l = 1, \dots, k$. Then, the value

$$Z_{in}^0 = \zeta_i(t_n) \quad (6.14)$$

yields a $\mathcal{O}(h^{k+1})$ approximation to the solution of (6.1). This interpolation procedure was first proposed by In't Hout (1992) for the numerical solution of delay differential equations. For the iterative solution of the nonlinear Runge–Kutta equations (6.1), the starting approximation (6.14) is proposed and analyzed by Calvo (2002).

The implementation of this approach is very simple. Using Newton's interpolation formula we have

$$Z_{in}^0 = Z_{i,n-1} + \nabla Z_{i,n-1} + \dots + \nabla^{k-1} Z_{i,n-1} \quad (6.15)$$

with backward differences given by $\nabla Z_{i,n} = Z_{i,n} - Z_{i,n-1}$, $\nabla^2 Z_{i,n} = \nabla Z_{i,n} - \nabla Z_{i,n-1}$, etc.

Numerical Study of Starting Approximations. We consider the Kepler problem with eccentricity $e = 0.6$ and initial values such that the period is 2π . With many different step sizes $h = 2\pi/N$ we compute $N+1$ steps with the Gauss method of order $p = 2s$ ($p = 4, 8, 12$). In the last step we compute the different starting approximations and their error $(\sum_{i=1}^s \|Z_{in} - Z_{in}^0\|^2)^{1/2}$ as a function of the step size h . The result is plotted in Fig. 6.1. There, the pictures also contain the global errors after one period. They allow us to localize the values of h , which are of practical interest.

We observe that the equistage approximation (6.15) also behaves like $\mathcal{O}(h^{k+1})$ when $k+1$ is larger than the order of the integrator. However, due to the increasing error constants, the accuracy is improved only for small step sizes. An optimal k could be estimated by checking the decrease of the backward differences $\|\nabla^j Z_{i,n-1}\|$. The error of the starting approximation obtained from the continuous output behaves like $\mathcal{O}(h^{s+1})$ (for the Gauss methods) and, in contrast to the equistage approximation, improves with increasing order. The approximations (6.7) of order $s+1$ and $s+2$ are a clear improvement. As a conclusion we find that for this example the equistage approximation (which is free from additional function evaluations) is preferable only for $s = 2$ (order 4). For higher order, the approximation

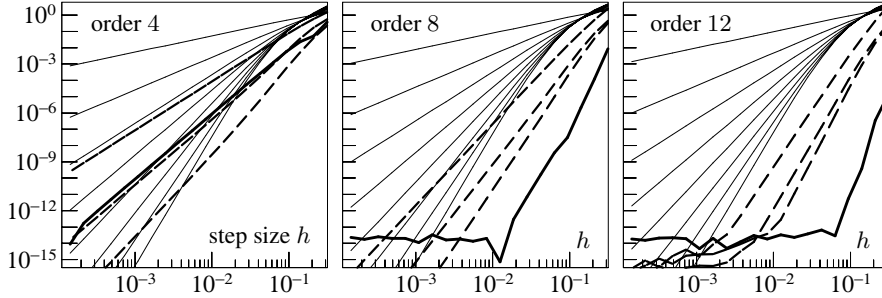


Fig. 6.1. Errors of starting approximations for Gauss methods as functions of the step size h : thick dashed lines for the extrapolated continuous output (6.4) and for the approximations (6.7) of order $s + 1$ and $s + 2$; thin solid lines for the equistage approximation (6.15) with $k = 0, 1, \dots, 7$; the thick solid line represents the global error of the method after one period

obtained from (6.7) is significantly more accurate and so it is worthwhile to spend these two additional function evaluations per step.

VIII.6.2 Fixed-Point Versus Newton Iteration

Finally we investigate the iterative solution of the nonlinear Runge–Kutta system (6.1). We discuss fixed-point and Newton-like iterations, and we compare their efficiency to the use of composition methods.

Fixed-Point Iteration. This is the most simple and most natural iteration for the solution of (6.1). With any starting approximation Z_{in}^0 from Sect. VIII.6.1 it reads

$$Z_{in}^{k+1} = h \sum_{j=1}^s a_{ij} f(y_n + Z_{jn}^k), \quad i = 1, \dots, s. \quad (6.16)$$

In the case where the entries of the Jacobian matrix $f'(y)$ are not excessively large (nonstiff problems) and that the step size is sufficiently small, this iteration converges for $k \rightarrow \infty$ to the solution of (6.1). Usually, the iteration is stopped if a certain norm of the differences $Z_{in}^{k+1} - Z_{in}^k$ is sufficiently small. We then use Z_{in}^k in the update formula (6.2) so that no additional function evaluation is required.

For a numerical study of the convergence of this iteration, we consider the Kepler problem with eccentricity $e = 0.6$ and initial values as in the preceding experiments (period of the solution is 2π). We apply the Gauss methods of order 4, 8, and 12 with various step sizes. For the integration over one period we show in Table 6.1 the total number of function evaluations, the mean number of required iterations per step, and the global error at the endpoint of integration. As a stopping criterion for the fixed-point iteration we check whether the norm of the difference of two successive approximations is smaller than 10^{-16} (roundoff unit in double precision). As a starting approximation Z_{in}^0 we use (6.15) with $k = 8$ for the method of order 4,

Table 6.1. Statistics of Gauss methods (total number of function evaluations, number of fixed-point iterations per step, and the global error at the endpoint) for computations of the Kepler problem over one period with $e = 0.6$

Fixed-point iteration (general problems)					
Gauss	$h = 2\pi/25$	$h = 2\pi/50$	$h = 2\pi/100$	$h = 2\pi/200$	$h = 2\pi/400$
order 4	803	1 043	1 393	1 825	2 319
	16.1	10.4	7.0	4.6	2.9
	$9.2 \cdot 10^{-2}$	$1.7 \cdot 10^{-2}$	$1.3 \cdot 10^{-3}$	$8.4 \cdot 10^{-5}$	$5.3 \cdot 10^{-6}$
order 8	1 021	1 455	2 091	3 007	4 183
	9.7	6.8	4.7	3.3	2.1
	$1.1 \cdot 10^{-3}$	$6.9 \cdot 10^{-7}$	$3.6 \cdot 10^{-9}$	$1.8 \cdot 10^{-11}$	$6.9 \cdot 10^{-14}$
order 12	1 297	1 731	2 311	3 441	5 917
	8.3	5.4	3.5	2.5	2.1
	$2.7 \cdot 10^{-6}$	$8.0 \cdot 10^{-11}$	$2.7 \cdot 10^{-14}$	$\leq \text{roundoff}$	$\leq \text{roundoff}$

and the approximation (6.7) of order $s + 2$ for the methods of orders 8 and 12. The coefficients are those presented after equation (6.12).

Since the starting approximations are more accurate for small h , the number of necessary iterations decreases drastically. In particular, for the 4th order method we need about 16 iterations per step for $h = 2\pi/25$, but at most 2 iterations when $h \leq 2\pi/800$. If one is interested in high accuracy computations (e.g., long-time simulations in astronomy), for which the error over one period is not larger than 10^{-10} , Table 6.1 illustrates that high order methods ($p \geq 12$) are most efficient.

Newton-Type Iterations. A standard technique for solving nonlinear equations is Newton's method or some modification of it. Writing the nonlinear system (6.1) of an implicit Runge–Kutta method as $F(Z) = 0$ with $Z = (Z_{1n}, \dots, Z_{sn})^T$, the Newton iteration is

$$Z^{k+1} = Z^k - M^{-1}F(Z^k), \quad (6.17)$$

where M is some approximation to the Jacobian matrix $F'(Z^k)$. Since the solution Z of the nonlinear system is $\mathcal{O}(h)$ close to zero, it is common to use $M = F'(0)$ so that the matrix M is independent of the iteration index k . In our special situation we get

$$M = I \otimes I - hA \otimes J \quad (6.18)$$

with $J = f'(y_n)$. Here, I denotes the identity matrix of suitable dimension, and A is the Runge–Kutta matrix.

We repeat the experiment of Table 6.1 with modified Newton iterations instead of fixed-point iterations. The result is shown in Table 6.2. We have suppressed the error at the end of the period, because it is the same as in Table 6.1. As expected, the convergence is faster (i.e., the number of iterations per step is smaller) so that the total number of function evaluations is reduced. However, we do not see in this table that we computed at every step the Jacobian $f'(y_n)$ and an LR -decomposition of the matrix M . Even if we exploit the tensor product structure in (6.18) as explained

Table 6.2. Statistics of Gauss methods (total number of function evaluations, number of iterations per step) for computations of the Kepler problem over one period with $e = 0.6$

Modified Newton iteration (general problems)					
Gauss	$h = 2\pi/25$	$h = 2\pi/50$	$h = 2\pi/100$	$h = 2\pi/200$	$h = 2\pi/400$
order 4	383	511	765	1 125	1 677
	7.7	5.1	3.8	2.8	2.1
order 8	597	883	1 387	2 307	3 667
	5.5	3.9	3.0	2.4	1.8
order 12	763	1 095	1 717	3 003	5 689
	4.7	3.3	2.5	2.2	2.0

in Hairer & Wanner (1996, Sect. IV.8), the cpu time is now considerably larger. Further improvements are possible, if the Jacobian of f and hence also the LR -decomposition of M is frozen over a couple of steps. But all these efforts can hardly beat (in cpu time) the straightforward fixed-point iterations. In accordance with the experience of Sanz-Serna & Calvo (1994, Sect. 5.5) we recommend in general the use of fixed-point iterations.

Separable Systems and Second Order Differential Equations. Many interesting differential equations are of the form

$$\dot{\eta} = f(y), \quad \dot{y} = g(\eta). \quad (6.19)$$

For example, the second order differential equation $\ddot{y} = f(y)$ is obtained by putting $g(\eta) = \eta$. Also Hamiltonian systems with separable Hamiltonian $H(p, q) = T(p) + U(q)$ are of the form (6.19).

For this particular system the Runge–Kutta equations (6.1) become

$$\zeta_{in} - h \sum_{j=1}^s a_{ij} f(y_n + Z_{jn}) = 0, \quad Z_{in} - h \sum_{j=1}^s a_{ij} g(\eta_n + \zeta_{jn}) = 0.$$

In this case we can still do better: instead of the standard fixed-point iteration (6.16) we apply a Gauss–Seidel like iteration

$$\zeta_{in}^{k+1} = h \sum_{j=1}^s a_{ij} f(y_n + Z_{jn}^k), \quad Z_{in}^{k+1} = h \sum_{j=1}^s a_{ij} g(\eta_n + \zeta_{jn}^{k+1}), \quad (6.20)$$

which is explicit for separable systems (6.19). Notice that the starting approximations have to be computed only for ζ_{in} . Those for Z_{in} are then obtained by (6.20) with $k + 1 = 0$.

For second order differential equations $\ddot{y} = f(y)$, where $g(\eta) = \eta$, this iteration becomes

$$Z_{in}^{k+1} = hc_i \eta_n + h^2 \sum_{j=1}^s \hat{a}_{ij} f(y_n + Z_{jn}^k), \quad (6.21)$$

Table 6.3. Statistics of iterations (6.20) for Gauss methods (total number of function evaluations, number of iterations per step) for computations of the Kepler problem over one period with $e = 0.6$

Fixed-point iteration (separable problems)					
Gauss	$h = 2\pi/25$	$h = 2\pi/50$	$h = 2\pi/100$	$h = 2\pi/200$	$h = 2\pi/400$
order 4	437	603	857	1 201	1 717
	8.7	6.0	4.3	3.0	2.1
order 8	613	923	1 427	2 339	3 647
	5.6	4.1	3.1	2.4	1.8
order 12	781	1 131	1 741	3 027	5 677
	4.9	3.4	2.6	2.2	2.0

where $c_i = \sum_{j=1}^s a_{ij}$ and \hat{a}_{ij} are the entries of the square A^2 of the Runge–Kutta matrix (any Nyström method could be applied as well). Due to the factor h^2 in (6.21) we expect this iteration to converge about twice as fast as the standard fixed-point iteration.

The Kepler problem is a second order differential equation, so that the iteration (6.21) can be applied. In analogy to the previous tables we present in Table 6.3 the statistics of such an implementation of the Gauss methods. We observe that for relatively large step sizes the number of iterations required per step is nearly halved (compared to Table 6.1). For high accuracy requirements the number of necessary iterations is surprisingly small, and the question arises whether such an implementation can compete with high order explicit composition methods.

Comparison Between Implicit Runge–Kutta and Composition Methods. We consider second order differential equations $\ddot{y} = f(y)$, so that composition methods based on the explicit Störmer–Verlet scheme can be applied. We use the coefficients of method (V.3.14) which has turned out to be excellent in the experiments of Sect. V.3.2. It is a method of order 8 and uses 17 function evaluations per integration step.

We compare it with the Gauss methods of order 8 and 12 (i.e., $s = 4$ and $s = 6$). As a starting approximation for the solution of the nonlinear system (6.1) we use (6.7) with $m = 3$, $\mu_1 = 0$, $\mu_2 = 1$, $\mu_3 = 1.75$, μ_{ij} chosen such that (6.11) holds for $k = 1, \dots, s + i - 1$, and β_{ij}, ν_{ij} such that order $s + 2$ is obtained. Since we are concerned with second order differential equations, we apply the iterations (6.20) until the norm of the difference of two successive approximations is below 10^{-17} .

For both classes of methods we use compensated summation (Algorithm 5.1), which permits us to reduce rounding errors. For composition methods we apply this technique for all updates of the basic integrator. For Runge–Kutta methods, we use it for adding the increment to y_n and also for computing the sum $\sum_{i=1}^s b_i k_i$.

The work–precision diagrams of the comparison are given in Fig. 6.2. The upper pictures correspond to the Kepler problem with $e = 0.6$ and an integration over 100 periods; the lower pictures correspond to the outer solar system with data given in Sect. I.2.4 and an integration over 500 000 earth days. The left pictures show the

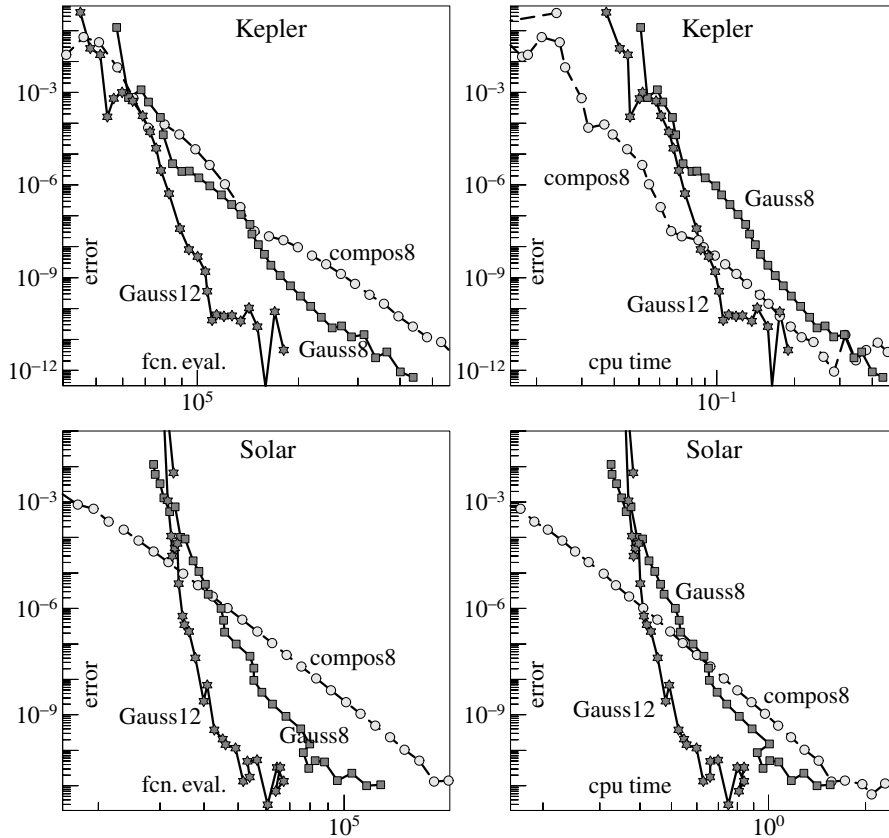


Fig. 6.2. Work–precision diagrams for two problems (Kepler and outer solar system) and three numerical integrators (composition method with coefficients of method (V.3.14) based on the explicit Störmer–Verlet scheme and the Gauss methods of orders 8 and 12)

Euclidean norm of the error at the end of the integration interval as a function of total numbers of function evaluations required for the integration; the pictures to the right present the same error as a function of the cpu times (with optimizing compiler on a SunBlade 100 workstation). We can draw the following conclusions from this experiment:

- the implementation of composition methods based on the Störmer–Verlet scheme is extremely easy; that of implicit Runge–Kutta methods is slightly more involved because it requires a stopping criterion for the fixed-point iterations;
- the overhead (total cpu time minus that used for the function evaluations) is much higher for the implicit Runge–Kutta methods; this is seen from the fact that implicit Runge–Kutta methods require less function evaluations for a given accuracy, but often more cpu time;
- among the two Gauss methods, the higher order method is more efficient for all precisions of practical interest;

- for very accurate computations (say, in quadruple precision), high order Runge–Kutta methods are more efficient than composition methods;
- much of the computation in the Runge–Kutta code can be done in parallel (e.g., the s function evaluations of a fixed-point iteration); composition methods do not have this potential;
- implicit Runge–Kutta methods can be applied to general (non-separable) differential equations, and the cost of the implementation is at most twice as large; if one is obliged to use an implicit method as the basic method for composition, many advantages of composition methods are lost.

Both classes of methods (composition and implicit Runge–Kutta) are of interest in the geometric integration of differential equations. Each one has its advantages and disadvantages.

Fortran codes of these computations are available on the Internet under the homepage <http://www.unige.ch/math/folks/haier/>. A Matlab version of these codes is described in E. & M. Hairer (2003).

VIII.7 Exercises

1. Consider a one-step method applied to a Hamiltonian system. Give a probabilistic proof of the property that the error of the numerical Hamiltonian due to roundoff grows like $\mathcal{O}(\sqrt{t} \, eps)$.
2. Prove that the collocation polynomial can be written as

$$w_n(t) = y_n + h \sum_{i=1}^s \beta_i(t) f(Y_{in}),$$

where the polynomials $\beta_i(t)$ are a solution of

$$\sum_{j=1}^s \beta_j(t) c_j^{k-1} = \frac{t^k}{k}.$$

3. Apply your favourite code to the Kepler problem and to the outer solar system with data as in Fig. 6.2. Plot a work-precision diagram.

Remark. Figure 7.1 shows our results obtained with the 8th order Runge–Kutta code Dop853 (Hairer, Nørsett & Wanner 1993) compared to an 8th order composition method. Rounding errors are more pronounced for Dop853, because compensated summation is not applied. Computations on shorter time intervals and comparisons of required function evaluations would be more in favour for Dop853. It is also of interest to consider high order Runge–Kutta Nyström methods.

4. Consider starting approximations

$$Y_{in}^0 = y_{n-2} + h \sum_{j=1}^s \beta_{ij}^{(2)} f(Y_{j,n-2}) + h \sum_{j=1}^s \beta_{ij}^{(1)} f(Y_{j,n-1}) \quad (7.1)$$

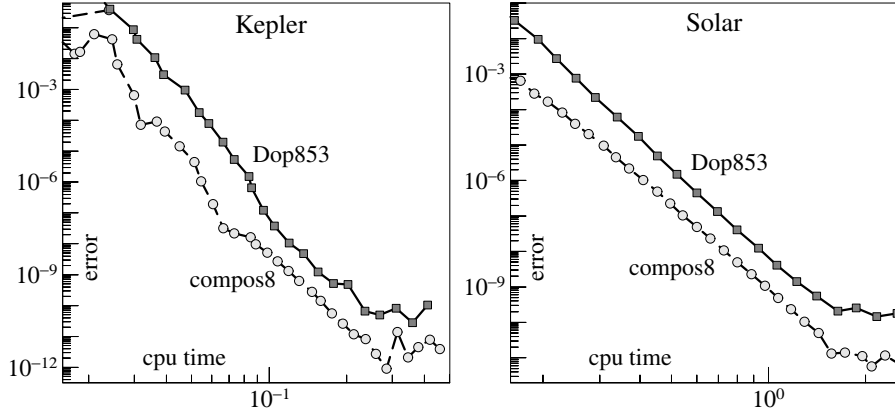


Fig. 7.1. Work–precision diagrams for the explicit, variable step size Runge–Kutta code Dop853 applied to two problems (Kepler and outer solar system). For a comparison, the results of Fig. 6.2 for the composition method are included

which use the internal stages of two consecutive steps without any additional function evaluation. What are the conditions such that (7.1) is of order $s + 1$, of order $s + 2$?

Compare the efficiency of these formulas with the algorithms (A) and (B) of Sect. VIII.6.1.

5. Prove that for a second order differential equation $\ddot{y} = f(y)$ (more precisely, for $\dot{y} = z, \dot{z} = f(y)$) the application of the s -stage Gauss method gives

$$y_{n+1} = y_n + h\dot{y}_n + h^2 \sum_{i=1}^s b_i(1 - c_i)f(y_n + Z_{in})$$

$$\dot{y}_{n+1} = \dot{y}_n + h \sum_{i=1}^s b_i f(y_n + Z_{in}),$$

where Z_{in} is obtained from the iteration (6.21).

Hint. The coefficients of the Gauss methods satisfy $\sum_j b_j a_{ji} = b_i(1 - c_i)$ for all i .

Chapter IX.

Backward Error Analysis and Structure Preservation

One of the greatest virtues of backward analysis ... is that when it is the appropriate form of analysis it tends to be very markedly superior to forward analysis. Invariably in such cases it has remarkable formal simplicity and gives deep insight into the stability (or lack of it) of the algorithm. (J.H. Wilkinson, IMA Bulletin 1986)

The origin of backward error analysis dates back to the work of Wilkinson (1960) in numerical linear algebra. For the study of integration methods for ordinary differential equations, its importance was seen much later. The present chapter is devoted to this theory. It is very useful, when the qualitative behaviour of numerical methods is of interest, and when statements over very long time intervals are needed. The formal analysis (construction of the modified equation, study of its properties) gives already a lot of insight into numerical methods. For a rigorous treatment, the modified equation, which is a formal series in powers of the step size, has to be truncated. The error, induced by such a truncation, can be made exponentially small, and the results remain valid on exponentially long time intervals.

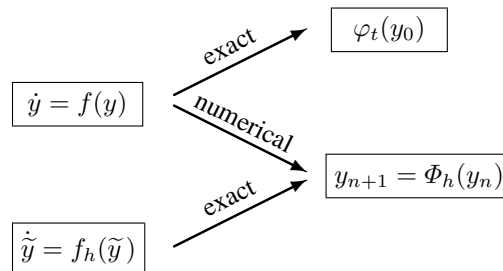
IX.1 Modified Differential Equation – Examples

Consider an ordinary differential equation

$$\dot{y} = f(y),$$

and a numerical method $\Phi_h(y)$ which produces the approximations

$$y_0, y_1, y_2, \dots$$



A forward error analysis consists of the study of the errors $y_1 - \varphi_h(y_0)$ (local error) and $y_n - \varphi_{nh}(y_0)$ (global error) in the solution space. The idea of backward error analysis is to search for a *modified differential equation* $\dot{\tilde{y}} = f_h(\tilde{y})$ of the form

$$\dot{\tilde{y}} = f(\tilde{y}) + hf_2(\tilde{y}) + h^2f_3(\tilde{y}) + \dots, \quad (1.1)$$

such that $y_n = \tilde{y}(nh)$, and in studying the difference of the vector fields $f(y)$ and $f_h(y)$. This then gives much insight into the qualitative behaviour of the numerical solution and into the global error $y_n - y(nh) = \tilde{y}(nh) - y(nh)$. We remark that the series in (1.1) usually diverges and that one has to truncate it suitably. The effect of such a truncation will be studied in Sect. IX.7. For the moment we content ourselves with a formal analysis without taking care of convergence issues. The idea of interpreting the numerical solution as the exact solution of a modified equation is common to many numerical analysts (“... This is possible since the map is the solution of some physical Hamiltonian problem which, in some sense, is close to the original problem”, Ruth (1983), or “... the symplectic integrator creates a numerical Hamiltonian system that is close to the original ...”, Gladman, Duncan & Candy 1991). A systematic study started with the work of Griffiths & Sanz-Serna (1986), Feng (1991), Sanz-Serna (1992), Yoshida (1993), Eirola (1993), Fiedler & Scheurle (1996), and many others.

For the computation of the modified equation (1.1) we put $y := \tilde{y}(t)$ for a fixed t , and we expand the solution of (1.1) into a Taylor series

$$\begin{aligned} \tilde{y}(t+h) &= y + h(f(y) + hf_2(y) + h^2 f_3(y) + \dots) \\ &\quad + \frac{h^2}{2!}(f'(y) + hf'_2(y) + \dots)(f(y) + hf_2(y) + \dots) + \dots \end{aligned} \quad (1.2)$$

We assume that the numerical method $\Phi_h(y)$ can be expanded as

$$\Phi_h(y) = y + hf(y) + h^2 d_2(y) + h^3 d_3(y) + \dots \quad (1.3)$$

(the coefficient of h is $f(y)$ for consistent methods). The functions $d_j(y)$ are known and are typically composed of $f(y)$ and its derivatives. For the explicit Euler method we simply have $d_j(y) = 0$ for all $j \geq 2$. In order to get $\tilde{y}(nh) = y_n$ for all n , we must have $\tilde{y}(t+h) = \Phi_h(y)$. Comparing like powers of h in the expressions (1.2) and (1.3) yields recurrence relations for the functions $f_j(y)$, namely,

$$\begin{aligned} f_2(y) &= d_2(y) - \frac{1}{2!}f'(y)f(y) \\ f_3(y) &= d_3(y) - \frac{1}{3!}(f''(f, f)(y) + f'f'f(y)) - \frac{1}{2!}(f'f_2(y) + f_2'f(y)). \end{aligned} \quad (1.4)$$

Example 1.1. Consider the scalar differential equation

$$\dot{y} = y^2, \quad y(0) = 1 \quad (1.5)$$

with exact solution $y(t) = 1/(1-t)$. It has a singularity at $t = 1$. We apply the explicit Euler method $y_{n+1} = y_n + hf(y_n)$ with step size $h = 0.02$. The picture in Fig. 1.1 presents the exact solution (dashed curve) together with the numerical solution (bullets). The above procedure for the computation of the modified equation, implemented as a Maple program (see Hairer & Lubich 2000) gives

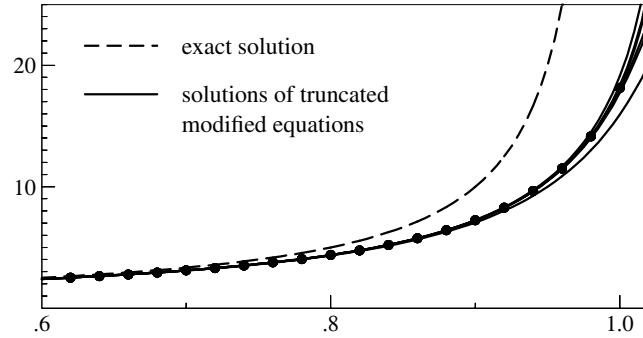


Fig. 1.1. Solutions of the modified equation for the problem (1.5)

```

> fcn := y -> y^2:
> nn := 6:
> fcoe[1] := fcn(y):
> for n from 2 by 1 to nn do
>   modeq := sum(h^j*fcoe[j+1], j=0..n-2):
>   diffy[0] := y:
>   for i from 1 by 1 to n do
>     diffy[i] := diff(diffy[i-1], y)*modeq:
>   od:
>   ytilde := sum(h^k*diffy[k]/k!, k=0..n):
>   res := ytilde-y-h*fcn(y):
>   tay := convert(series(res, h=0, n+1), polynom):
>   fcoe[n] := -coeff(tay, h, n):
> od:
> simplify(sum(h^j*fcoe[j+1], j=0..nn-1));

```

Its output is

$$\dot{\tilde{y}} = \tilde{y}^2 - h\tilde{y}^3 + h^2 \frac{3}{2}\tilde{y}^4 - h^3 \frac{8}{3}\tilde{y}^5 + h^4 \frac{31}{6}\tilde{y}^6 - h^5 \frac{157}{15}\tilde{y}^7 \pm \dots \quad (1.6)$$

The above picture also presents the solution of the modified equation, when truncated after 1, 2, 3, and 4 terms. We observe an excellent agreement of the numerical solution with the exact solution of the modified equation.

A similar program for the implicit midpoint rule (I.1.7) computes the modified equation

$$\dot{\tilde{y}} = \tilde{y}^2 + h^2 \frac{1}{4}\tilde{y}^4 + h^4 \frac{1}{8}\tilde{y}^6 + h^6 \frac{11}{192}\tilde{y}^8 + h^8 \frac{3}{128}\tilde{y}^{10} \pm \dots, \quad (1.7)$$

and for the classical Runge–Kutta method of order 4 (left tableau of (II.1.8))

$$\dot{\tilde{y}} = \tilde{y}^2 - h^4 \frac{1}{24}\tilde{y}^6 + h^6 \frac{65}{576}\tilde{y}^8 - h^7 \frac{17}{96}\tilde{y}^9 + h^8 \frac{19}{144}\tilde{y}^{10} \pm \dots \quad (1.8)$$

We observe that the perturbation terms in the modified equation are of size $\mathcal{O}(h^p)$, where p is the order of the method. This is true in general.

Theorem 1.2. Suppose that the method $y_{n+1} = \Phi_h(y_n)$ is of order p , i.e.,

$$\Phi_h(y) = \varphi_h(y) + h^{p+1}\delta_{p+1}(y) + \mathcal{O}(h^{p+2}),$$

where $\varphi_t(y)$ denotes the exact flow of $\dot{y} = f(y)$, and $h^{p+1}\delta_{p+1}(y)$ the leading term of the local truncation error. The modified equation then satisfies

$$\tilde{y}' = f(\tilde{y}) + h^p f_{p+1}(\tilde{y}) + h^{p+1} f_{p+2}(\tilde{y}) + \dots, \quad \tilde{y}(0) = y_0 \quad (1.9)$$

with $f_{p+1}(y) = \delta_{p+1}(y)$.

Proof. The construction of the functions $f_j(y)$ (see the beginning of this section) shows that $f_j(y) = 0$ for $2 \leq j \leq p$ if and only if $\Phi_h(y) - \varphi_h(y) = \mathcal{O}(h^{p+1})$. \square

A first application of the modified equation (1.1) is the existence of an *asymptotic expansion of the global error*. Indeed, by the nonlinear variation of constants formula, the difference between its solution $\tilde{y}(t)$ and the solution $y(t)$ of $\dot{y} = f(y)$ satisfies

$$\tilde{y}(t) - y(t) = h^p e_p(t) + h^{p+1} e_{p+1}(t) + \dots \quad (1.10)$$

Since $y_n = \tilde{y}(nh) + \mathcal{O}(h^N)$ for the solution of a truncated modified equation, this proves the existence of an asymptotic expansion in powers of h for the global error $y_n - y(nh)$.

A large part of this chapter studies properties of the modified differential equation, and the question of the extent to which structures (such as conservation of invariants, Hamiltonian structure) in the problem $\dot{y} = f(y)$ can carry over to the modified equation.

Example 1.3. We next consider the Lotka–Volterra equations

$$\dot{q} = q(p-1), \quad \dot{p} = p(2-q),$$

and we apply (a) the explicit Euler method, and (b) the symplectic Euler method, both with constant step size $h = 0.1$. The first terms of their modified equations are

$$\begin{aligned} \text{(a)} \quad \dot{q} &= q(p-1) - \frac{h}{2} q(p^2 - pq + 1) + \mathcal{O}(h^2), \\ \dot{p} &= -p(q-2) - \frac{h}{2} p(q^2 - pq - 3q + 4) + \mathcal{O}(h^2), \\ \text{(b)} \quad \dot{q} &= q(p-1) - \frac{h}{2} q(p^2 + pq - 4p + 1) + \mathcal{O}(h^2), \\ \dot{p} &= -p(q-2) + \frac{h}{2} p(q^2 + pq - 5q + 4) + \mathcal{O}(h^2). \end{aligned}$$

Figure 1.2 shows the numerical solutions for initial values indicated by a thick dot. In the pictures to the left they are embedded in the exact flow of the differential equation, whereas in those to the right they are embedded in the flow of the modified differential equation, truncated after the h^2 terms. As in the first example, we observe an excellent agreement of the numerical solution with the exact solution of

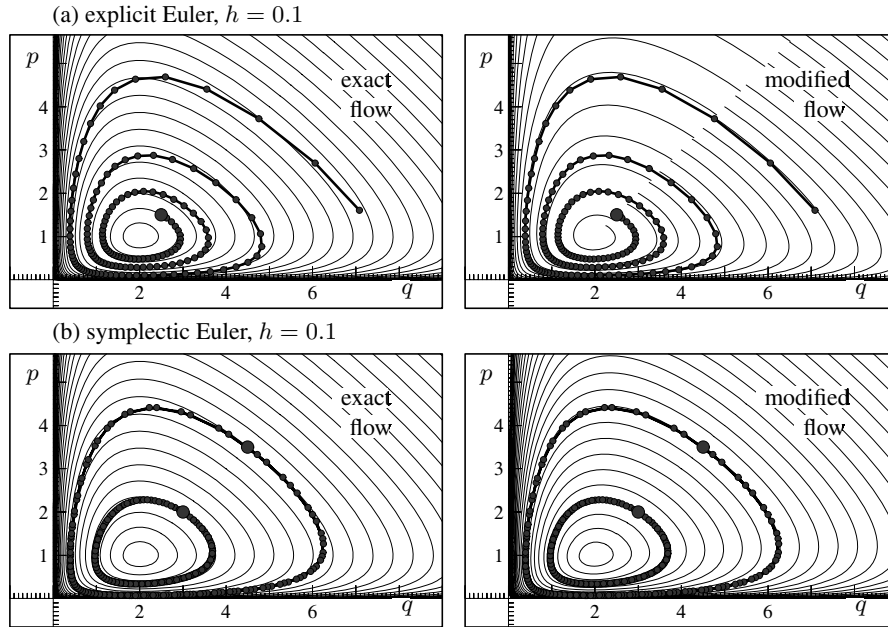


Fig. 1.2. Numerical solution compared to the exact and modified flows

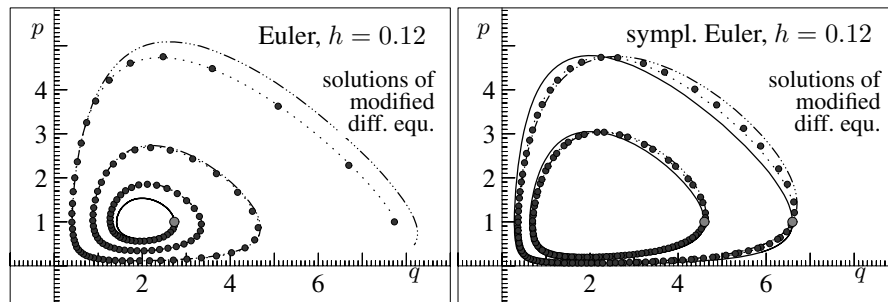


Fig. 1.3. Study of the truncation in the modified equation

the modified equation. For the symplectic Euler method, the solutions of the truncated modified equation are periodic, as is the case for the unperturbed problem (Exercise 5).

In Fig. 1.3 we present the numerical solution and the exact solution of the modified equation, once truncated after the h terms (dashed-dotted), and once truncated after the h^2 terms (dotted). The exact solution of the problem is included as a solid curve. This shows that taking more terms in the modified equation usually improves the agreement of its solution with the numerical approximation of the method.

Example 1.4. For a linear differential equation with constant coefficients

$$\dot{y} = Ay, \quad y(0) = y_0$$

we consider numerical methods which yield $y_{n+1} = R(hA)y_n$, where $R(z)$ is the stability function (VI.4.9) of the method. In this case we get $y_n = R(hA)^n y_0$, so that $y_n = \tilde{y}(nh)$, where $\tilde{y}(t) = R(hA)^{t/h} y_0 = \exp\left(\frac{t}{h} \ln R(hA)\right) y_0$ is the solution of the modified differential equation

$$\dot{\tilde{y}} = \frac{1}{h} \ln R(hA) \tilde{y} = (A + hb_2A^2 + h^2b_3A^3 + \dots) \tilde{y} \quad (1.11)$$

with suitable constants b_2, b_3, \dots . Since $R(z) = 1 + z + \mathcal{O}(z^2)$ and $\ln(1+x) = x - x^2/2 + \mathcal{O}(x^3)$ both have a positive radius of convergence, the series (1.11) converges for $|h| < h_0$ with some $h_0 > 0$. We shall see later that this is an exceptional situation. In general, the modified equation is a formal divergent series.

IX.2 Modified Equations of Symmetric Methods

In this and the following sections we investigate how the structure of the differential equation and geometric properties of the method are reflected in the modified differential equation. Here we begin by studying this question for symmetric/reversible methods.

Consider a numerical method Φ_h . Recall that its adjoint $y_{n+1} = \Phi_h^*(y_n)$ is defined by the relation $y_n = \Phi_{-h}(y_{n+1})$ (see Definition II.1.4).

Theorem 2.1 (Adjoint Methods). *Let $f_j(y)$ be the coefficient functions of the modified equation for the method Φ_h . Then, the coefficient functions $f_j^*(y)$ of the modified equation for the adjoint method Φ_h^* satisfy*

$$f_j^*(y) = (-1)^{j+1} f_j(y). \quad (2.1)$$

Proof. The solution $\tilde{y}(t)$ of the modified equation for Φ_h^* has to satisfy $\tilde{y}(t) = \Phi_{-h}(\tilde{y}(t+h))$ or, equivalently, $\tilde{y}(t-h) = \Phi_{-h}(y)$ with $y := \tilde{y}(t)$. We get (2.1) if we replace h with $-h$ in the formulas (1.1), (1.2) and (1.3). \square

For symmetric methods we have $\Phi_h^* = \Phi_h$, implying $f_j^*(y) = f_j(y)$. We therefore get the following corollary to Theorem 2.1.

Theorem 2.2 (Symmetric Methods). *The coefficient functions of the modified equation of a symmetric method satisfy $f_j(y) = 0$ whenever j is even, so that (1.1) has an expansion in even powers of h .* \square

This theorem explains the h^2 -expansion in the modified equation (1.7) of the midpoint rule.

As a consequence of Theorem 2.2, the asymptotic expansion (1.10) of the global error is also in even powers of h . This property is responsible for the success of h^2 -extrapolation methods.

Consider now a numerical method applied to a ρ -reversible differential equation as studied in Sect. V.1. Recall from Theorem V.1.5 that a symmetric method is ρ -reversible under the ρ -compatibility condition (V.1.4), which is satisfied for most numerical methods.

Theorem 2.3 (Reversible Methods). *Consider a ρ -reversible differential equation $\dot{y} = f(y)$ and a ρ -reversible numerical method $\Phi_h(y)$. Then, every truncation of the modified differential equation is again ρ -reversible.*

Proof. Let $f_j(y)$ be the j th coefficient of the modified equation (1.1) for Φ_h . The proof is by induction on j . So assume that for $j = 1, \dots, r$, the vector field $f_j(y)$ is ρ -reversible, i.e.,

$$\rho \circ f_j = -f_j \circ \rho.$$

We show that the same relation holds also for $j = r + 1$. By assumption, the truncated modified equation

$$\dot{\tilde{y}} = f(\tilde{y}) + hf_2(\tilde{y}) + \dots + h^{r-1}f_r(\tilde{y})$$

is ρ -reversible, so that by (V.1.2), it has a ρ -reversible flow $\varphi_{r,t}(y)$, that is, $\rho \circ \varphi_{r,t} = \varphi_{r,t}^{-1} \circ \rho$. By construction of the modified equation, we have

$$\Phi_h(y) = \varphi_{r,h}(y) + h^{r+1}f_{r+1}(y) + \mathcal{O}(h^{r+2}).$$

Since $\varphi_{r,h}(y) = y + \mathcal{O}(h)$, this implies

$$\Phi_h^{-1}(y) = \varphi_{r,h}^{-1}(y) - h^{r+1}f_{r+1}(y) + \mathcal{O}(h^{r+2}).$$

Since both Φ_h and $\varphi_{r,h}$ are ρ -reversible maps, these two relations yield $\rho \circ f_{r+1} = -f_{r+1} \circ \rho$ as desired. \square

IX.3 Modified Equations of Symplectic Methods

We now present one of the most important results of this chapter. We consider a Hamiltonian system $\dot{y} = J^{-1}\nabla H(y)$ with an infinitely differentiable Hamiltonian $H(y)$, and we show that the modified equation of symplectic methods is also Hamiltonian.

IX.3.1 Existence of a Local Modified Hamiltonian

... if we neglect convergence questions then one can always find a formal integral ... (J. Moser 1968)

Theorem 3.1. *If a symplectic method $\Phi_h(y)$ is applied to a Hamiltonian system with a smooth Hamiltonian $H : \mathbb{R}^{2d} \rightarrow \mathbb{R}$, then the modified equation (1.1) is also Hamiltonian. More precisely, there exist smooth functions $H_j : \mathbb{R}^{2d} \rightarrow \mathbb{R}$ for $j = 2, 3, \dots$, such that $f_j(y) = J^{-1}\nabla H_j(y)$.*

The following proof by induction, whose ideas can be traced back to Moser (1968), was given by Benettin & Giorgilli (1994) and Tang (1994). It can be extended to many other situations. We have already encountered its reversible version in the proof of Theorem 2.3.

Proof. Assume that $f_j(y) = J^{-1}\nabla H_j(y)$ for $j = 1, 2, \dots, r$ (this is satisfied for $r = 1$, because $f_1(y) = f(y) = J^{-1}\nabla H(y)$). We have to prove the existence of a Hamiltonian $H_{r+1}(y)$. The idea is to consider the truncated modified equation

$$\dot{\tilde{y}} = f(\tilde{y}) + hf_2(\tilde{y}) + \dots + h^{r-1}f_r(\tilde{y}), \quad (3.1)$$

which is a Hamiltonian system with Hamiltonian $H(y) + hH_2(y) + \dots + h^{r-1}H_r(y)$. Its flow $\varphi_{r,t}(y_0)$, compared to that of (1.1), satisfies

$$\Phi_h(y_0) = \varphi_{r,h}(y_0) + h^{r+1}f_{r+1}(y_0) + \mathcal{O}(h^{r+2}),$$

and also

$$\Phi'_h(y_0) = \varphi'_{r,h}(y_0) + h^{r+1}f'_{r+1}(y_0) + \mathcal{O}(h^{r+2}).$$

By our assumption on the method and by the induction hypothesis, Φ_h and $\varphi_{r,h}$ are symplectic transformations. This, together with $\varphi'_{r,h}(y_0) = I + \mathcal{O}(h)$, therefore implies

$$J = \Phi'_h(y_0)^T J \Phi'_h(y_0) = J + h^{r+1} \left(f'_{r+1}(y_0)^T J + J f'_{r+1}(y_0) \right) + \mathcal{O}(h^{r+2}).$$

Consequently, the matrix $J f'_{r+1}(y)$ is symmetric and the existence of $H_{r+1}(y)$ satisfying $f_{r+1}(y) = J^{-1}\nabla H_{r+1}(y)$ follows from the Integrability Lemma VI.2.7. This part of the proof is similar to that of Theorem VI.2.6. \square

For Hamiltonians $H : D \rightarrow \mathbb{R}$ the statement of the above theorem remains valid with $H_j : D \rightarrow \mathbb{R}$ on domains $D \subset \mathbb{R}^{2d}$ on which the Integrability Lemma VI.2.7 is applicable. This is the case for simply connected domains D , but not in general (see the discussion after the proof of Lemma VI.2.7).

IX.3.2 Existence of a Global Modified Hamiltonian

By Lemma VI.5.3 every symplectic one-step method $\Phi_h : (p, q) \mapsto (P, Q)$ can be locally expressed in terms of a generating function $S(P, q, h)$ as

$$p = P + \frac{\partial S}{\partial q}(P, q, h), \quad Q = q + \frac{\partial S}{\partial P}(P, q, h). \quad (3.2)$$

This property allows us to give an independent proof of Theorem 3.1 and in addition to show that the modified equation is Hamiltonian with $\tilde{H}(p, q)$ defined on the same domain as the generating function. The following result is mentioned in Benettin & Giorgilli (1994) and in the thesis of Murua (1994), p. 100.

Theorem 3.2. Assume that the symplectic method Φ_h has a generating function

$$S(P, q, h) = h S_1(P, q) + h^2 S_2(P, q) + h^3 S_3(P, q) + \dots \quad (3.3)$$

with smooth $S_j(P, q)$ defined on an open set D . Then, the modified differential equation is a Hamiltonian system with

$$\tilde{H}(p, q) = H(p, q) + h H_2(p, q) + h^2 H_3(p, q) + \dots, \quad (3.4)$$

where the functions $H_j(p, q)$ are defined and smooth on the whole of D .

Proof. By Theorem VI.5.7, the exact solution $(P, Q) = (\tilde{p}(t), \tilde{q}(t))$ of the Hamiltonian system corresponding to $\tilde{H}(p, q)$ is given by

$$p = P + \frac{\partial \tilde{S}}{\partial q}(P, q, t), \quad Q = q + \frac{\partial \tilde{S}}{\partial P}(P, q, t),$$

where \tilde{S} is the solution of the Hamilton–Jacobi differential equation

$$\frac{\partial \tilde{S}}{\partial t}(P, q, t) = \tilde{H}\left(P, q + \frac{\partial \tilde{S}}{\partial P}(P, q, t)\right), \quad \tilde{S}(P, q, 0) = 0. \quad (3.5)$$

Since \tilde{H} depends on the parameter h , this is also the case for \tilde{S} . Our aim is to determine the functions $H_j(p, q)$ such that the solution $\tilde{S}(P, q, t)$ of (3.5) coincides for $t = h$ with (3.3).

We first express $\tilde{S}(P, q, t)$ as a series

$$\tilde{S}(P, q, t) = t \tilde{S}_1(P, q, h) + t^2 \tilde{S}_2(P, q, h) + t^3 \tilde{S}_3(P, q, h) + \dots,$$

insert it into (3.5) and compare powers of t . This allows us to obtain the functions $\tilde{S}_j(p, q, h)$ recursively in terms of derivatives of \tilde{H} :

$$\begin{aligned} \tilde{S}_1(p, q, h) &= \tilde{H}(p, q) \\ 2 \tilde{S}_2(p, q, h) &= \left(\frac{\partial \tilde{H}}{\partial q} \cdot \frac{\partial \tilde{S}_1}{\partial P} \right)(p, q, h) \\ 3 \tilde{S}_3(p, q, h) &= \left(\frac{\partial \tilde{H}}{\partial q} \cdot \frac{\partial \tilde{S}_2}{\partial P} \right)(p, q, h) + \frac{1}{2} \left(\frac{\partial^2 \tilde{H}}{\partial q^2} \left(\frac{\partial \tilde{S}_1}{\partial P}, \frac{\partial \tilde{S}_1}{\partial P} \right) \right)(p, q, h). \end{aligned} \quad (3.6)$$

We then write \tilde{S}_j as a series

$$\tilde{S}_j(p, q, h) = \tilde{S}_{j1}(p, q) + h \tilde{S}_{j2}(p, q) + h^2 \tilde{S}_{j3}(p, q) + \dots,$$

insert it and the expansion (3.4) for \tilde{H} into (3.6), and compare powers of h . This yields $\tilde{S}_{1k}(p, q) = H_k(p, q)$ and for $j > 1$ we see that $\tilde{S}_{jk}(p, q)$ is a function of derivatives of H_l with $l < k$.

The requirement $S(p, q, h) = \tilde{S}(p, q, h)$ finally shows $S_1(p, q) = \tilde{S}_{11}(p, q)$, $S_2(p, q) = \tilde{S}_{12}(p, q) + \tilde{S}_{21}(p, q)$, etc., so that

$$S_j(p, q) = H_j(p, q) + \text{“function of derivatives of } H_k(p, q) \text{ with } k < j\text{”}.$$

For a given generating function $S(P, q, h)$, this recurrence relation allows us to determine successively the $H_j(p, q)$. We see from these explicit formulas that the functions H_j are defined on the same domain as the S_j . \square

As a consequence of Theorem 3.2 and Theorems VI.5.4 and VI.5.5 we obtain the following result.

Theorem 3.3. *A symplectic (partitioned) Runge–Kutta method applied to a system with smooth Hamiltonian $H : D \rightarrow \mathbb{R}$ (with $D \subset \mathbb{R}^{2d}$ an arbitrary open set) has a modified Hamiltonian (3.4) with smooth functions $H_j : D \rightarrow \mathbb{R}$.* \square

Example 3.4 (Symplectic Euler Method). The symplectic Euler method is nothing other than (3.2) with $S(P, q, h) = h H(P, q)$. We therefore have (3.3) with $S_1(p, q) = H(p, q)$ and $S_j(p, q) = 0$ for $j > 1$. Following the constructive proof of Theorem 3.2 we obtain

$$\tilde{H} = H - \frac{h}{2} H_p H_q + \frac{h^2}{12} (H_{pp} H_q^2 + H_{qq} H_p^2 + 4H_{pq} H_q H_p) + \dots \quad (3.7)$$

as the modified Hamiltonian of the symplectic Euler method. For vector-valued p and q , the expression $H_p H_q$ is the scalar product of the vectors H_p and H_q , and $H_{pp} H_q^2 = H_{pp}(H_q, H_q)$ with the second derivative interpreted as a bilinear mapping.

As a particular example consider the pendulum problem (I.1.13), which is Hamiltonian with $H(p, q) = p^2/2 - \cos q$, and apply the symplectic Euler method. By (3.7), the modified Hamiltonian is

$$\tilde{H}(p, q) = H(p, q) - \frac{h}{2} p \sin q + \frac{h^2}{12} (\sin^2 q + p^2 \cos q) + \dots$$

This example illustrates that the modified equation corresponding to a separable Hamiltonian (i.e., $H(p, q) = T(p) + U(q)$) is in general not separable. Moreover, it shows that the modified equation of a second order differential equation $\ddot{q} = -\nabla U(q)$ (or equivalently, $\dot{q} = p, \dot{p} = -\nabla U(q)$) is in general not a second order equation.

In principle, the constructive proof of Theorem 3.2 allows us to explicitly compute the modified equation of every symplectic (partitioned) Runge–Kutta method. In Sect. IX.9.3 below we shall, however, give explicit formulas for the modified Hamiltonian in terms of trees. This also yields an alternative proof of Theorem 3.3.

IX.3.3 Poisson Integrators

Consider a Poisson system, i.e., a differential equation

$$\dot{y} = B(y)\nabla H(y), \quad (3.8)$$

where the structure matrix $B(y)$ satisfies the conditions of Lemma VII.2.3, and apply a Poisson integrator (Definition VII.4.6).

Theorem 3.5. *If a Poisson integrator $\Phi_h(y)$ is applied to the Poisson system (3.8), then the modified equation is locally a Poisson system. More precisely, for every $y_0 \in \mathbb{R}^n$ there exist a neighbourhood U and smooth functions $H_j : U \rightarrow \mathbb{R}$ such that on U , the modified equation is of the form*

$$\dot{\tilde{y}} = B(\tilde{y})\left(\nabla H(\tilde{y}) + h \nabla H_2(\tilde{y}) + h^2 \nabla H_3(\tilde{y}) + \dots\right). \quad (3.9)$$

Proof. We use the local change of coordinates $(u, c) = \chi(y)$ of the Darboux–Lie Theorem. By Corollary VII.3.6, this transforms (3.8) to

$$\dot{u} = J^{-1} \nabla_u K(u, c), \quad \dot{c} = 0,$$

where $K(u, c) = H(y)$ and ∇_u is the gradient with respect to u . The same transformation takes $\Phi_h(y)$ to $\chi \circ \Phi_h \circ \chi^{-1}(u, c) = (\Psi_h^1(u, c), c)$, where by Lemma VII.4.10 $u \mapsto \Psi_h^1(u, c)$ is a symplectic transformation for every c . By Theorem 3.1, the modified equation in the (u, c) variables is of the form

$$\dot{\tilde{u}} = J^{-1} \nabla_u \tilde{K}(\tilde{u}, \tilde{c}), \quad \dot{\tilde{c}} = 0$$

with $\tilde{K}(u, c) = K(u, c) + h K_2(u, c) + h^2 K_3(u, c) + \dots$. Transforming back to the y -variables gives the modified equation (3.9) with $H_j(y) = K_j(u, c)$. \square

The above result is purely local in that it relies on the local transformation of the Darboux–Lie Theorem. It can be made more global under additional conditions on the differential equation.

Theorem 3.6. *If $H(y)$ and $B(y)$ are defined and smooth on a simply connected domain D , and if $B(y)$ is invertible on D , then a Poisson integrator $\Phi_h(y)$ has a modified equation (3.9) with smooth functions $H_j(y)$ defined on all of D .*

Proof. By the construction of Sect. IX.1, the coefficient functions $f_j(y)$ of the modified equation (1.1) are defined and smooth on D . Since $B(y)$ is assumed invertible, there exist unique smooth functions $g_j(y)$ such that $f_j(y) = B(y)g_j(y)$. It remains to show that $g_j(y) = \nabla H_j(y)$ for a function $H_j(y)$ defined on D .

By the local result of Theorem 3.5, we know that for every $y_0 \in D$ there exist functions $H_j^0(y)$ such that $g_j(y) = \nabla H_j^0(y)$ in a neighbourhood of y_0 . This implies that the Jacobian of $g_j(y)$ is symmetric on D . The Integrability Lemma VI.2.7 thus proves the existence of functions $H_j(y)$ defined on all of D such that $g_j(y) = \nabla H_j(y)$. \square

IX.4 Modified Equations of Splitting Methods

For splitting methods applied to a differential equation

$$\dot{y} = f^{[1]}(y) + f^{[2]}(y), \quad (4.1)$$

the modified differential equation is obtained directly with the calculus of Lie derivatives and the Baker-Campbell-Hausdorff formula. This approach is due to Yoshida (1993) who considered the case of separable Hamiltonian systems.

First-Order Splitting. Consider the splitting method

$$\Phi_h = \varphi_h^{[1]} \circ \varphi_h^{[2]},$$

where $\varphi_h^{[i]}$ is the time- h flow of $\dot{y} = f^{[i]}(y)$. In terms of the Lie derivatives D_i defined by $D_i g(y) = g'(y) f^{[i]}(y)$, this method becomes, using Lemma III.5.1,

$$\Phi_h = \exp(hD_2) \exp(hD_1) \text{Id},$$

and with the BCH formula (III.4.11), (III.4.12) this reads

$$\Phi_h = \exp(h\tilde{D}) \text{Id}$$

with

$$\tilde{D} = D_1 + D_2 + \frac{h}{2} [D_2, D_1] + \frac{h^2}{12} ([D_2, [D_2, D_1]] + [D_1, [D_1, D_2]]) + \dots \quad (4.2)$$

It follows that Φ_h is formally the exact time- h flow of the modified equation

$$\tilde{y}' = \tilde{f}(\tilde{y}) \quad \text{with} \quad \tilde{f} = \tilde{D} \text{Id}. \quad (4.3)$$

This gives

$$\tilde{f}(y) = f(y) + hf_2(y) + h^2 f_3(y) + \dots$$

with $f = f^{[1]} + f^{[2]}$ and

$$\begin{aligned} f_2 &= \frac{1}{2} (f^{[1]'} f^{[2]} - f^{[2]'} f^{[1]}) \\ f_3 &= \frac{1}{12} (f^{[1]''} (f^{[2]}, f^{[2]}) + f^{[1]'} f^{[2]'} f^{[2]} - f^{[2]''} (f^{[1]}, f^{[2]}) - f^{[2]'} f^{[1]'} f^{[2]} \\ &\quad + f^{[2]''} (f^{[1]}, f^{[1]}) + f^{[2]'} f^{[1]'} f^{[1]} - f^{[1]''} (f^{[2]}, f^{[1]}) - f^{[1]'} f^{[2]'} f^{[1]}). \end{aligned}$$

Strang Splitting. For the symmetric splitting

$$\Phi_h^{[S]} = \varphi_{h/2}^{[1]} \circ \varphi_h^{[2]} \circ \varphi_{h/2}^{[1]}$$

the symmetric BCH formula (III.4.14), (III.4.15) yields

$$\Phi_h^{[S]} = \exp\left(\frac{h}{2}D_1\right) \exp(hD_2) \exp\left(\frac{h}{2}D_1\right) \text{Id} = \exp(h\tilde{D}^{[S]}) \text{Id}$$

with

$$\tilde{D}^{[S]} = D_1 + D_2 + h^2 \left(-\frac{1}{24}[D_1, [D_1, D_2]] + \frac{1}{12}[D_2, [D_2, D_1]] \right) + \dots \quad (4.4)$$

Hence, $\Phi_h^{[S]}$ is the formally exact flow of the modified equation

$$\dot{\tilde{y}} = \tilde{f}^{[S]}(\tilde{y}) \quad \text{with} \quad \tilde{f}^{[S]} = \tilde{D}^{[S]} \text{Id}. \quad (4.5)$$

This gives

$$\tilde{f}^{[S]}(y) = f(y) + h^2 f_3^{[S]}(y) + h^4 f_5^{[S]}(y) + \dots$$

with $f = f^{[1]} + f^{[2]}$ and

$$\begin{aligned} f_3^{[S]} = & \left(\frac{1}{12} (f^{[1]''}(f^{[2]}, f^{[2]}) + f^{[1]'} f^{[2]'} f^{[2]} - f^{[2]''}(f^{[1]}, f^{[2]}) - f^{[2]'} f^{[1]'} f^{[2]}) \right. \\ & \left. - \frac{1}{24} (f^{[2]''}(f^{[1]}, f^{[1]}) + f^{[2]'} f^{[1]'} f^{[1]} - f^{[1]''}(f^{[2]}, f^{[1]}) - f^{[1]'} f^{[2]'} f^{[1]}) \right). \end{aligned}$$

The modified equations for general splitting methods (III.5.13) are obtained in the same way, using Lemma III.5.5.

Hamiltonian Splittings. Consider a differential equation (4.1) where the vector fields $f^{[i]}(y) = J^{-1} \nabla H^{[i]}(y)$ are Hamiltonian. Lemma VII.3.1 shows that the commutator of the Lie derivatives of two Hamiltonian vector fields is the Lie derivative of another Hamiltonian vector field which corresponds to the Poisson bracket of the two Hamiltonians: $[D_F, D_G] = D_{\{G, F\}}$. This implies in particular that the modified differential equations (4.3) and (4.5) are again Hamiltonian. For the first-order splitting, we thus get $f_j(y) = J^{-1} \nabla H_j(y)$, where by (4.2) and (4.3),

$$\begin{aligned} H_2 &= \frac{1}{2} \{H^{[1]}, H^{[2]}\} \\ H_3 &= \frac{1}{12} \left(\{ \{H^{[1]}, H^{[2]}\}, H^{[2]} \} + \{ \{H^{[2]}, H^{[1]}\}, H^{[1]} \} \right), \end{aligned}$$

and for the Strang splitting, by (4.4) and (4.5),

$$H_3^{[S]} = -\frac{1}{24} \{ \{H^{[2]}, H^{[1]}\}, H^{[1]} \} + \frac{1}{12} \{ \{H^{[1]}, H^{[2]}\}, H^{[2]} \}.$$

The explicit expressions from the BCH-formula show that the modified Hamiltonian is defined on the same open set as the smooth Hamiltonians $H^{[i]}$.

For the splitting $H(p, q) = T(p) + U(q)$ of a separable Hamiltonian, this approach gives an alternative derivation of the modified equation (3.7) of the symplectic Euler method, and a simple construction of the modified equation of the Störmer–Verlet method (Yoshida 1993). Here, the formula simplifies to

$$\tilde{H}^{[S]} = H + h^2 \left(-\frac{1}{24} U_{qq}(T_p, T_p) + \frac{1}{12} T_{pp}(U_q, U_q) \right) + \dots \quad (4.6)$$

IX.5 Modified Equations of Methods on Manifolds

We consider the relationship between numerical methods for differential equations on manifolds and the associated modified differential equations. We give applications to the study of first integrals, constrained Hamiltonian systems, and Lie–Poisson integrators.

IX.5.1 Methods on Manifolds and First Integrals

Consider a differential equation on a smooth manifold \mathcal{M} ,

$$\dot{y} = f(y) \quad \text{with} \quad f(y) \in T_y \mathcal{M}, \quad (5.1)$$

with a smooth vector field $f(y)$ defined on \mathcal{M} .

Theorem 5.1. *Let $\Phi_h : \mathcal{M} \rightarrow \mathcal{M}$ be an integrator on the manifold \mathcal{M} , with $\Phi_h(y)$ depending smoothly on (y, h) . Then, there exists a modified differential equation on \mathcal{M} ,*

$$\dot{\tilde{y}} = f(\tilde{y}) + h f_2(\tilde{y}) + h^2 f_3(\tilde{y}) + \dots \quad (5.2)$$

with smooth $f_j(y) \in T_y \mathcal{M}$, such that $\varphi_{r,h}(y) = \Phi_h(y) + \mathcal{O}(h^{r+1})$, where $\varphi_{r,t}(y)$ denotes the flow of the truncation of (5.2) after r terms.

For symmetric methods, the expansion (5.2) contains only even powers of h .

Proof. We choose a local parametrization $y = \chi(z)$ of the manifold \mathcal{M} . In the coordinates z the differential equation (5.1) reads

$$\dot{z} = F(z) \quad \text{with } F(z) \text{ defined by} \quad \chi'(z)F(z) = f(\chi(z)),$$

and the numerical integrator becomes

$$\Psi_h(z) = \chi^{-1} \circ \Phi_h \circ \chi(z).$$

Since $F(z)$ and $\Psi_h(z)$ are smooth, the standard backward error analysis on \mathbb{R}^n of Sect. IX.1 yields a modified equation for the integrator $\Psi_h(z)$,

$$\dot{\tilde{z}} = F(\tilde{z}) + h F_2(\tilde{z}) + h^2 F_3(\tilde{z}) + \dots$$

Defining

$$f_j(y) = \chi'(z) F_j(z) \quad \text{for} \quad y = \chi(z)$$

gives the desired vector fields $f_j(y)$ on \mathcal{M} . It follows from the uniqueness of the modified equation in the parameter space that $f_j(y)$ is independent of the choice of the local parametrization.

The additional statement on symmetric methods follows from Theorem 2.2, because Ψ_h is symmetric if and only if Φ_h is symmetric. \square

Under an analyticity assumption, the converse statement also holds.

Theorem 5.2. *Let the integrator $\Phi_h : U \rightarrow \mathbb{R}^n$ (with open $U \subset \mathbb{R}^n$) be real analytic in h , and let $\mathcal{M} = \{y \in U; g(y) = 0\}$ with real analytic $g : U \rightarrow \mathbb{R}^m$. If the coefficient functions $f_j(y)$ of the modified differential equation (5.2) satisfy $g'(y)f_j(y) = 0$ for all j and all $y \in \mathcal{M}$, then the restriction of Φ_h to \mathcal{M} defines an integrator on \mathcal{M} , i.e., $\Phi_h : \mathcal{M} \rightarrow \mathcal{M}$.*

Proof. By the assumption on $f_j(y)$, the flow of the truncated modified equation satisfies $g \circ \varphi_{r,h}(y) = 0$ for all $r \geq 1$ and all $y \in \mathcal{M}$. Since $\varphi_{r,h}(y) = \Phi_h(y) + \mathcal{O}(h^{r+1})$, we have $g \circ \Phi_h(y) = \mathcal{O}(h^{r+1})$ for all r . The analyticity assumptions therefore imply $g \circ \Phi_h(y) = 0$. \square

Theorems 5.1 and 5.2 apply to many situations treated in Chap. IV.

First Integrals. The following result was obtained by Gonzalez, Higham & Stuart (1999) and Reich (1999) with different arguments.

Corollary 5.3. *Consider a differential equation $\dot{y} = f(y)$ with a first integral $I(y)$, i.e., $I'(y)f(y) = 0$ for all y . If the numerical method preserves this first integral, then every truncation of the modified equation has $I(y)$ as a first integral.*

Proof. This follows from Theorem 5.1 by considering $\dot{y} = f(y)$ as a differential equation on the manifold $\mathcal{M} = \{y; I(y) = \text{Const}\}$, for which the tangent space is $T_y\mathcal{M} = \{v; I'(y)v = 0\}$. \square

The following converse of Corollary 5.3 is a direct consequence of Theorem 5.2.

Corollary 5.4. *Consider a differential equation $\dot{y} = f(y)$ with a real-analytic first integral $I(y)$. If the numerical method $\Phi_h(y)$ is real analytic in h , and if every truncation of the modified equation has $I(y)$ as a first integral, then the numerical method preserves $I(y)$ exactly, i.e., $I(\Phi_h(y)) = I(y)$ for all y .* \square

Projection Methods. Algorithm IV.4.2 defines a smooth mapping on the manifold if the direction of projection depends smoothly on the position. This is satisfied by orthogonal projection, but is not fulfilled if switching coordinate projections are used (as in Example 4.3). The symmetric orthogonal projection method of Algorithm V.4.1 gives a symmetric method on the manifold to which Theorem 5.1 can be applied.

Methods Based on Local Coordinates. If the parametrization of the manifold employed in Algorithms IV.5.3 and V.4.5 depends smoothly on the position, then again Theorem 5.1 applies. This is the case for the tangent space parametrization, but not for the generalized coordinate partitioning considered in Sect. IV.5.3.

Corollary 5.5 (Lie Group Methods). *Consider a differential equation on a matrix Lie group G ,*

$$\dot{Y} = A(Y)Y,$$

where $A(Y)$ is in the associated Lie algebra \mathfrak{g} . A Lie group integrator $\Phi_h : G \rightarrow G$ has the modified equation

$$\dot{\tilde{Y}} = (A(\tilde{Y}) + hA_2(\tilde{Y}) + h^2A_3(\tilde{Y}) + \dots)\tilde{Y} \quad (5.3)$$

with $A_j(Y) \in \mathfrak{g}$ for $Y \in G$.

Proof. This is a direct consequence of Theorem 5.1 and (IV.6.3), viz., $T_Y G = \{AY | A \in \mathfrak{g}\}$. \square

IX.5.2 Constrained Hamiltonian Systems

In Sect. VII.1 we studied symplectic numerical integrators for constrained Hamiltonian systems

$$\begin{aligned} \dot{q} &= H_p(p, q) \\ \dot{p} &= -H_q(p, q) - G(q)^T \lambda \\ 0 &= g(q). \end{aligned} \quad (5.4)$$

Assuming the regularity condition (VII.1.13), the Lagrange parameter $\lambda = \lambda(p, q)$ is given by (VII.1.12). This system can be interpreted as a differential equation on the manifold

$$\mathcal{M} = \{(p, q) \mid g(q) = 0, G(q)H_p(p, q) = 0\}, \quad (5.5)$$

where $G(q) = g'(q)$. The symplectic Euler method (VII.1.19)–(VII.1.20), the RAT-TLE scheme (VII.1.26), and the Lobatto IIIA-IIIIB pair (VII.1.27)–(VII.1.30) were found to be symplectic integrators Φ_h on the manifold \mathcal{M} .

Theorem 5.6. *A symplectic integrator $\Phi_h : \mathcal{M} \rightarrow \mathcal{M}$ for the constrained Hamiltonian system (5.4) has a modified equation which is locally of the form*

$$\begin{aligned} \dot{\tilde{q}} &= \tilde{H}_p(\tilde{p}, \tilde{q}) \\ \dot{\tilde{p}} &= -\tilde{H}_q(\tilde{p}, \tilde{q}) - G(\tilde{q})^T \tilde{\lambda} \\ 0 &= g(\tilde{q}), \end{aligned} \quad (5.6)$$

where $\tilde{\lambda} = \tilde{\lambda}(\tilde{p}, \tilde{q})$ is given by (VII.1.12) with H replaced by \tilde{H} , and

$$\tilde{H}(p, q) = H(p, q) + h H_2(p, q) + h^2 H_3(p, q) + \dots \quad (5.7)$$

with $H_j(p, q)$ satisfying $G(q)\nabla_p H_j(p, q) = 0$ for $(p, q) \in \mathcal{M}$ and all j .

Proof. As explained in Example VII.2.7, a local parametrization $(p, q) = \chi(z)$ of the manifold \mathcal{M} transforms (5.4) to the Poisson system

$$\dot{z} = B(z)\nabla K(z) \quad (5.8)$$

with $B(z) = (\chi'(z)^T J \chi'(z))^{-1}$ and $K(z) = H(\chi(z))$. Lemma VII.4.9 implies that the numerical method $\Phi_h(p, q)$ on \mathcal{M} becomes a Poisson integrator $\Psi_h(z)$ for (5.8). By Theorem 3.5, $\Psi_h(z)$ has the modified equation

$$\dot{\tilde{z}} = B(\tilde{z})\left(\nabla K(\tilde{z}) + h \nabla K_2(\tilde{z}) + h^2 \nabla K_3(\tilde{z}) + \dots\right). \quad (5.9)$$

Let π be a smooth projection onto the manifold \mathcal{M} , defined on a neighbourhood of \mathcal{M} in \mathbb{R}^{2d} . We then define

$$H_j(p, q) = K_j(\chi^{-1}(\pi(p, q))) + \mu(p, q)^T G(q) \nabla_p H(p, q)$$

where we choose $\mu(p, q)$ such that

$$G(q) \nabla_p H_j(p, q) = 0 \quad \text{for } (p, q) \in \mathcal{M}. \quad (5.10)$$

This is possible because of the regularity assumption (VII.1.13), and because $G(q) \nabla_p H(p, q) = 0$ on \mathcal{M} . The condition (5.10) implies that the system (5.6) can be viewed as a differential equation on the original manifold \mathcal{M} . Using the same parametrization $(p, q) = \chi(z)$ as before shows that (5.6) is equivalent to (5.9). \square

We note that, due to the arbitrary choice of the projection π , the functions $H_j(p, q)$ of the modified equation are uniquely defined only on \mathcal{M} .

Global Modified Hamiltonian. If we restrict our considerations to partitioned Runge–Kutta methods, it is possible to find $H_j(p, q)$ in (5.7) that are globally defined on \mathcal{M} . Such a result is proved by Reich (1996a) and by Hairer & Wanner (1996) for the constrained symplectic Euler method and the RATTLE algorithm, and by Hairer (2003) for general symplectic partitioned Runge–Kutta schemes. We follow the approach of the latter publication, but present the result only for the important special case of the RATTLE algorithm (VII.1.26). The construction of the $H_j(p, q)$ is done in the following three steps.

Step 1. Symplectic Extension of the Method to a Neighbourhood of the Manifold. The numerical solution (p_1, q_1) of (VII.1.26) is well-defined only for initial values satisfying $(p_0, q_0) \in \mathcal{M}$. However, if we replace the condition “ $g(q_1) = 0$ ” by

$$g(q_1) = g(q_0) + h G(q_0) H_p(p_0, q_0), \quad (5.11)$$

and the condition “ $G(q_1) H_p(p_1, q_1) = 0$ ” by

$$G(q_1) H_p(p_1, q_1) = G(q_0) H_p(p_0, q_0), \quad (5.12)$$

then the numerical solution is well-defined for all (p_0, q_0) in an h -independent open neighbourhood of \mathcal{M} (cf. the existence and uniqueness proof of Sect. VII.1.3). Unfortunately, the so-obtained extension of (VII.1.26) is not symplectic.

Inspired by the formula of Lasagni for the generating function of (unconstrained) symplectic Runge–Kutta methods (see Sect. VI.5.2), we let

$$\begin{aligned} S(p_1, q_0, h) &= \frac{h}{2} \left(H(p_{1/2}, q_0) + H(p_{1/2}, q_1) + g(q_0)^T \lambda + g(q_1)^T \mu \right) \\ &\quad - \frac{h^2}{4} \left(H_q(p_{1/2}, q_1) + G(q_1)^T \mu \right)^T \left(H_p(p_{1/2}, q_0) + H_p(p_{1/2}, q_1) \right), \end{aligned} \quad (5.13)$$

where $p_0, p_{1/2}, p_1, q_0, q_1, \lambda, \mu$ are the values of the above extension. In the definition (5.13) of the generating function we consider $p_0, p_{1/2}, q_1, \lambda, \mu$ as functions of

(p_1, q_0) , what is possible because $p_1 = p_0 + \mathcal{O}(h)$. With the help of $S(p, q, h)$ we define a new numerical method on a neighbourhood of \mathcal{M} by

$$p_0 = p_1 + S_q(p_1, q_0, h), \quad q_1 = q_0 + S_p(p_1, q_0, h). \quad (5.14)$$

This method is symplectic by definition, and it also coincides with the RATTLE algorithm on the manifold \mathcal{M} . Using the fact that the last expression in (5.13) equals $(p_1 - p_{1/2})^T(q_1 - q_0)$, this is seen by the same computation as in the proof of Theorem VI.5.4.

Step 2. Application of the Results of Sect. IX.3.2. The function $S(p_1, q_0, h)$ of (5.13) can be expanded into powers of h with coefficients depending on (p_1, q_0) . These coefficient functions are composed of derivatives of $H(p, q)$ and $g(q)$ and, consequently, they are globally defined. For example, the h -coefficient is

$$S_1(p_1, q_0) = H(p_1, q_0) + g(q_0)^T \lambda(p_1, q_0), \quad (5.15)$$

where $\lambda(p, q)$ is the function defined in (VII.1.12).

We are thus exactly in the situation, where we can apply Theorem 3.2. This proves that the method (5.14) has a modified differential equation with globally defined modified Hamiltonian

$$\tilde{H}_{ext}(p, q) = H_1(p, q) + hH_2(p, q) + \dots \quad (5.16)$$

In particular, the constructive proof of Theorem 3.2 shows that $H_1(p, q) = S_1(p, q)$ with $S_1(p, q)$ from (5.15).

Step 3. Backinterpretation for the Method on the Manifold. Since the RATTLE algorithm defines a one-step method on \mathcal{M} , it follows from Theorem 5.1 that every truncation of the modified differential equation

$$\dot{\tilde{p}} = -\nabla_q \tilde{H}_{ext}(\tilde{p}, \tilde{q}), \quad \dot{\tilde{q}} = \nabla_p \tilde{H}_{ext}(\tilde{p}, \tilde{q}) \quad (5.17)$$

is a differential equation on the manifold \mathcal{M} . Terms of the form $g(q)^T \mu(p, q)$ in $\tilde{H}_{ext}(p, q)$, which vanish on \mathcal{M} , give rise to $-g(q)^T \mu_q(p, q) - G(q)^T \mu(p, q)$ and $g(q)^T \mu_p(p, q)$ in the vector field of (5.17). On the manifold \mathcal{M} , where $g(q) = 0$, only the expression $-G(q)^T \mu(p, q)$ remains. Consequently, we can arbitrarily remove terms of the form $g(q)^T \mu(p, q)$ from the functions $H_j(p, q)$ in (5.16), if we add a term $-G(q)^T \lambda$ in the differential equation for p with λ defined by the relation $g(q) = 0$. This then gives a problem of the form (5.6) with globally defined $H_j(p, q)$.

IX.5.3 Lie–Poisson Integrators

As in Sect. VII.5.5 we consider a symplectic integrator

$$(P_1, Q_1) = \Phi_h(P_0, Q_0) \quad \text{on } T^*G$$

for the left-invariant Hamiltonian system (VII.5.43) on a matrix Lie group G with a Hamiltonian $H(P, Q)$ that is quadratic in P . We suppose that the method preserves the left-invariance (VII.5.54) so that it induces a one-step map

$$Y_1 = \Psi_h(Y_0) \quad \text{on } \mathfrak{g}^*$$

by setting $Y_1 = Q_1^T P_1$ for $(P_1, Q_1) = \Phi_h(P_0, Q_0)$ with $Q_0^T P_0 = Y_0$. This is a numerical integrator for the differential equation (VII.5.37) on \mathfrak{g}^* , and in the coordinates $y = (y_j)$ with respect to the basis (F_j) of \mathfrak{g}^* this gives a map

$$y_1 = \psi_h(y_0) \quad \text{on } \mathbb{R}^d,$$

which is a numerical integrator for the Lie–Poisson system $\dot{y} = B(y)\nabla H(y)$ with $B(y)$ given by (VII.5.35).

Theorem 5.7. *If $\Phi_h(P, Q)$ is a symplectic and left-invariant integrator for (VII.5.43) which is real analytic in h , then its reduction $\psi_h(y)$ is a Poisson integrator. Moreover, $\Psi_h(Y)$ preserves the coadjoint orbits, i.e., $\Psi_h(Y) \in \{\text{Ad}_{U^{-1}}^* Y; U \in G\}$.*

Proof. (a) In the first step one shows, by the standard induction argument as in the proof of Theorem 2.3, that the modified equation given by Theorem 5.6,

$$\begin{aligned} \dot{\tilde{P}} &= -\nabla_Q \tilde{H}(\tilde{P}, \tilde{Q}) - \sum_{i=1}^m \tilde{\lambda}_i \nabla_Q g_i(\tilde{Q}), & \dot{\tilde{Q}} &= \nabla_P \tilde{H}(\tilde{P}, \tilde{Q}) \\ 0 &= g_i(\tilde{Q}), \quad i = 1, \dots, m, \end{aligned} \quad (5.18)$$

with

$$\tilde{H}(P, Q) = H(P, Q) + hH_2(P, Q) + h^2H_3(P, Q) + \dots$$

is left-invariant, i.e.,

$$H_j(U^T P, U^{-1} Q) = H_j(P, Q) \quad \text{for all } U \in G \text{ and all } j. \quad (5.19)$$

(b) The Lie–Poisson reduction of Theorem VII.5.8 yields that if $(\tilde{P}(t), \tilde{Q}(t)) \in T^*G$ is a solution of the modified system (5.18), then $\tilde{Y}(t) = \tilde{Q}(t)^T \tilde{P}(t) \in \mathfrak{g}^*$ solves the differential equation

$$\langle \dot{\tilde{Y}}, X \rangle = \langle \tilde{Y}, [\tilde{H}'(\tilde{Y}), X] \rangle \quad \text{for all } X \in \mathfrak{g}. \quad (5.20)$$

Theorem VII.5.6 shows that its solution lies on a coadjoint orbit. By Theorem VII.5.5, (5.20) is equivalent to the Poisson system

$$\dot{\tilde{y}} = B(\tilde{y})\nabla \tilde{H}(\tilde{y}). \quad (5.21)$$

(c) We know already from Theorem VII.5.11 that $\psi_h(y)$ is a Poisson map. Since all truncations of the modified equation (5.21) have the Casimirs as first integrals, their preservation by ψ_h follows from Corollary 5.4. Similarly, the preservation of the coadjoint orbits follows from Theorem 5.2. \square

In contrast to Theorem 3.5, we here obtain a global modified Hamiltonian in the modified Poisson system if the method is obtained by the discrete Lie–Poisson reduction of the RATTLE algorithm; see the preceding subsection.

IX.6 Modified Equations for Variable Step Sizes

The modified differential equation of a numerical integrator depends on the step size employed. Therefore, if the step size is changed arbitrarily, a different modified equation occurs at every step. This is the reason for the poor longtime behaviour observed in Sect. VIII.1. On the other hand, a satisfactory backward error analysis is possible for the variable-step approaches of Sects. VIII.2 and VIII.3.

Time Transformations. The adaptive approaches of Sect. VIII.2 amount to applying a fixed step size method to a transformed differential equation. Hence, the backward error analysis considered so far applies directly and yields modified equations for the transformed problem. These modified equations are Hamiltonian for Algorithm VIII.2.1 and reversible for method (VIII.2.12).

Proportional, Reversible Step Size Controllers. As in Sect. VIII.3.1 we let the step size be of the form

$$h_{n+1/2} = \varepsilon s(y_n, \varepsilon), \quad (6.1)$$

where ε is a small accuracy parameter. It is not allowed to use information from previous steps. The idea is to work with expansions in powers of the fixed parameter ε instead of the step sizes, and to consider the exact solution of the modified equation on a variable grid. The following development is given in Hairer & Stoffer (1997). It extends the results of Sects. IX.1 and IX.2 to variable step sizes.

Theorem 6.1. *Let $\Phi_h(y)$ be a smooth one-step method.*

a) The variable-step method $y \mapsto \Phi_{\varepsilon s(y, \varepsilon)}(y)$ has a modified differential equation

$$\dot{\tilde{y}} = f(\tilde{y}) + \varepsilon f_2(\tilde{y}) + \varepsilon^2 f_3(\tilde{y}) + \dots, \quad (6.2)$$

with smooth vector fields $f_j(y)$, such that

$$\varphi_{r, \varepsilon s(y, \varepsilon)}(y) = \Phi_{\varepsilon s(y, \varepsilon)}(y) + \mathcal{O}(\varepsilon^{r+1}), \quad (6.3)$$

where $\varphi_{r, t}(y)$ denotes the flow of the truncation of (6.2) after r terms.

b) If the method is symmetric (i.e., $\Phi_h(y) = \Phi_{-h}^{-1}(y)$) and $s(\hat{y}, -\varepsilon) = s(y, \varepsilon)$ holds with $\hat{y} = \Phi_{\varepsilon s(y, \varepsilon)}(y)$, then the expansion (6.2) is in even powers of ε , i.e.,

$$f_j(y) = 0 \quad \text{for even } j. \quad (6.4)$$

c) If the method is ρ -reversible (i.e., $\rho \circ \Phi_h = \Phi_h^{-1} \circ \rho$) and $s(\rho^{-1}\hat{y}, \varepsilon) = s(y, \varepsilon)$ holds with $\hat{y} = \Phi_{\varepsilon s(y, \varepsilon)}(y)$, then the modified equation (6.2) is ρ -reversible, i.e.,

$$\rho \circ f_j = -f_j \circ \rho \quad \text{for all } j. \quad (6.5)$$

Proof. a) The modified equation (6.2) is constructed by Taylor expansion of (6.3) in the same way as (1.1), using ε -expansions instead of h -expansions.

For the proof of the statements (b) and (c) we denote, as we did in Sect. VIII.3, $\Psi_\varepsilon(y) = \Phi_{\varepsilon s(y, \varepsilon)}(y)$. We then compute the dominant error term in (6.3) and obtain

$$\Psi_\varepsilon(y) = \varphi_{r,\varepsilon s(y,\varepsilon)}(y) + \varepsilon^{r+1}s(y,\varepsilon)f_{r+1}(y) + \mathcal{O}(\varepsilon^{r+2}). \quad (6.6)$$

With the aim of getting an analogous formula for Ψ_ε^{-1} , we put $\hat{y} = \Psi_\varepsilon(y)$ and use $\varphi_{r,t}^{-1}(y) = \varphi_{r,-t}(y)$ so that

$$y = \varphi_{r,-\varepsilon s(y,\varepsilon)}(\hat{y} - \varepsilon^{r+1}s(y,\varepsilon)f_{r+1}(y) + \mathcal{O}(\varepsilon^{r+2})). \quad (6.7)$$

b) Inserting $s(y,\varepsilon) = s(\hat{y}, -\varepsilon)$ into (6.7) and using the facts that $y = \hat{y} + \mathcal{O}(\varepsilon)$ and that the derivative $\varphi'_{r,t}(y)$ is $\mathcal{O}(t)$ -close to the identity, we obtain

$$\Psi_\varepsilon^{-1}(\hat{y}) = y = \varphi_{r,-\varepsilon s(\hat{y}, -\varepsilon)}(\hat{y}) - \varepsilon^{r+1}s(\hat{y}, 0)f_{r+1}(\hat{y}) + \mathcal{O}(\varepsilon^{r+2}). \quad (6.8)$$

By (VIII.3.3) we have $\Psi_\varepsilon = \Psi_{-\varepsilon}^{-1}$. Changing the sign of ε in (6.8), a comparison with (6.6) proves that $f_{r+1}(y) = (-1)^r f_{r+1}(y)$ implying (6.4).

c) With $s(y,\varepsilon) = s(\rho^{-1}\hat{y}, \varepsilon)$ formula (6.7) yields

$$\Psi_\varepsilon^{-1}(\hat{y}) = \varphi_{r,-\varepsilon s(\rho^{-1}\hat{y}, \varepsilon)}(\hat{y}) - \varepsilon^{r+1}s(\rho^{-1}\hat{y}, 0)f_{r+1}(\hat{y}) + \mathcal{O}(\varepsilon^{r+2}).$$

By an induction argument on r we assume that $\rho \circ \varphi_{r,t} = \varphi_{r,-t} \circ \rho$. The ρ -reversibility of Ψ_ε , i.e., $\rho \circ \Psi_\varepsilon = \Psi_\varepsilon^{-1} \circ \rho$, thus implies the statement (6.5). \square

Integrating, Reversible Step Size Controllers. We next study a backward error analysis for Algorithm VIII.3.4. It is possible to interpret this algorithm as the fixed step size method $\hat{\Phi}_\varepsilon$ of (VIII.3.19) applied to the augmented system (VIII.3.17) and to apply the construction of Sect. IX.1. This approach has been taken in Hairer & Söderlind (2004). In view of an error analysis for reversible integrable systems it seems to be more convenient to consider the solution of the modified equation on a variable grid as it is done in Theorem 6.1.

Let us recall Algorithm VIII.3.4. For a given basic integrator $\Phi_h(y)$ and a given time transformation $\sigma(y)$ we denote $G(y) = -(\sigma(y))^{-1} \nabla \sigma(y)^T f(y)$ and we compute for a given initial value y_0 and with $z_0 = 1/\sigma(y_0)$

$$\begin{aligned} z_{n+1/2} &= z_n + \varepsilon G(y_n)/2 \\ y_{n+1} &= \Phi_{\varepsilon/z_{n+1/2}}(y_n) \\ z_{n+1} &= z_{n+1/2} + \varepsilon G(y_{n+1})/2. \end{aligned} \quad (6.9)$$

The values y_n approximate $y(t_n)$, where $t_{n+1} = t_n + \varepsilon/z_{n+1/2}$. We further use the notation

$$\Psi_\varepsilon : \begin{pmatrix} y_n \\ z_n \end{pmatrix} \mapsto \begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} \quad \text{and} \quad \hat{\rho} = \begin{pmatrix} \rho & 0 \\ 0 & 1 \end{pmatrix}. \quad (6.10)$$

The step size used in this algorithm is

$$h_{n+1/2} = \frac{\varepsilon}{z_{n+1/2}} = \varepsilon s(y_n, z_n, \varepsilon) \quad \text{with} \quad s(y, z, \varepsilon) = \frac{1}{z + \varepsilon G(y)/2}. \quad (6.11)$$

The symmetric definition of the algorithm immediately yields

$$s(\hat{y}, \hat{z}, -\varepsilon) = s(y, z, \varepsilon) \quad \text{for} \quad (\hat{y}, \hat{z}) = \Psi_\varepsilon(y, z). \quad (6.12)$$

For a ρ -reversible differential equation $\dot{y} = f(y)$ and for $\sigma(y)$ satisfying $\sigma(\rho^{-1}y) = \sigma(y)$ we have $G(\rho^{-1}y) = -G(y)$. Consequently, the step size function $s(y, z, \varepsilon)$ of (6.11) also satisfies

$$s(\rho^{-1}\hat{y}, \hat{z}, -\varepsilon) = s(y, z, \varepsilon) \quad \text{for} \quad (\hat{y}, \hat{z}) = \Psi_\varepsilon(y, z). \quad (6.13)$$

With this preparation we are able to formulate the following result.

Theorem 6.2. *Let $\Phi_h(y)$ be a smooth one-step method, $\sigma(y)$ a smooth time transformation, and $s(y, z, \varepsilon)$ the step size function of (6.11).*

a) For the method Ψ_ε of (6.10) there exists a modified differential equation

$$\begin{aligned} \dot{\tilde{y}} &= f(\tilde{y}) + \varepsilon f_2(\tilde{y}, \tilde{z}) + \varepsilon^2 f_3(\tilde{y}, \tilde{z}) + \dots \\ \dot{\tilde{z}} &= \tilde{z} G(\tilde{y}) + \varepsilon G_2(\tilde{y}, \tilde{z}) + \varepsilon^2 G_3(\tilde{y}, \tilde{z}) + \dots, \end{aligned} \quad (6.14)$$

with smooth vector fields $f_j(y, z), G_j(y, z)$, such that

$$\varphi_{r,\varepsilon s(y,z,\varepsilon)}(y, z) = \Psi_\varepsilon(y, z) + \mathcal{O}(\varepsilon^{r+1}), \quad (6.15)$$

where $\varphi_{r,t}(y, z)$ denotes the flow of the truncation of the system (6.14) after r terms.

b) If the basic method is symmetric (i.e., $\Phi_h(y) = \Phi_{-h}^{-1}(y)$) then

$$f_j(y) = 0 \quad \text{for even } j. \quad (6.16)$$

c) If the basic method is ρ -reversible (i.e., $\rho \circ \Phi_h = \Phi_h^{-1} \circ \rho$) and $\sigma(\rho^{-1}y) = \sigma(y)$ holds, then the modified equation (6.14) is $\hat{\rho}$ -reversible with $\hat{\rho}$ given by (6.10), i.e.,

$$\rho f_j(y, z) = -f_j(\rho y, z), \quad G_j(y, z) = -G(\rho y, z) \quad \text{for all } j. \quad (6.17)$$

Proof. The proof is the same as for Theorem 6.1 and therefore omitted. Notice that the step size function satisfies (6.12) and (6.13) which are needed in that proof. \square

If the basic method is of order p then the coefficient functions of (6.14) satisfy $f_j(y, z) = 0$ for $j = 2, \dots, p$. We always have $G_2(y, z) = 0$ due to the symmetric way of choosing $z_{n+1/2}$ in (6.9). However, $G_3(y, z) \neq 0$ in general, even if the method Φ_h has an order higher than two.

IX.7 Rigorous Estimates – Local Error

Wherefore it is highly desirable that it be clearly and rigorously shown why series of this kind, which at first converge very rapidly and then ever more slowly, and at length diverge more and more, nevertheless give a sum close to the true one if not too many terms are taken, and to what degree such a sum can safely be considered as exact.

(a footnote in Gauss' thesis, 1799)

Up to now we have considered the modified equation (1.1) as a formal series without taking care of convergence issues. Here,

- we show that already in very simple situations the modified differential equation does not converge;
- we give bounds on the coefficient functions $f_j(y)$ of the modified equation (1.1), so that an optimal truncation index can be determined;
- we estimate the difference between the numerical solution $y_1 = \Phi_h(y_0)$ and the exact solution $\tilde{y}(h)$ of the truncated modified equation.

These estimates will be the basis for rigorous statements concerning the long-time behaviour of numerical solutions. The rigorous estimates of the present section have been given in the articles Benettin & Giorgilli (1994), Hairer & Lubich (1997) and Reich (1999). We mainly follow the approach of Benettin & Giorgilli, but we also use ideas of the other two papers.

Example 7.1. We consider the differential equation¹ $\dot{y} = f(t)$, $y(0) = 0$, and we apply the trapezoidal rule $y_1 = h(f(0) + f(h))/2$. In this case, the numerical solution has an expansion $\Phi_h(t, y) = y + h(f(t) + f(t+h))/2 = y + hf(t) + h^2 f'(t)/2 + h^3 f''(t)/4 + \dots$, so that the modified equation is necessarily of the form

$$\dot{\tilde{y}} = f(t) + hb_1 f'(t) + h^2 b_2 f''(t) + h^3 b_3 f'''(t) + \dots \quad (7.1)$$

The real coefficients b_k can be computed by putting $f(t) = e^t$. The relation $\Phi_h(t, y) = \tilde{y}(t+h)$ (with initial value $\tilde{y}(t) = y$) yields after division by e^t

$$\frac{h}{2}(e^h + 1) = (1 + b_1 h + b_2 h^2 + b_3 h^3 + \dots)(e^h - 1).$$

This proves that $b_1 = 0$, and $b_k = B_k/k!$, where B_k are the Bernoulli numbers (see for example Hairer & Wanner (1997), Sect. II.10). Since these numbers behave like $B_k/k! \approx \text{Const} \cdot (2\pi)^{-k}$ for $k \rightarrow \infty$, the series (7.1) diverges for all $h \neq 0$, as soon as the derivatives of $f(t)$ grow like $f^{(k)}(t) \approx k! MR^{-k}$. This is typically the case for analytic functions $f(t)$ with finite poles.

It is interesting to remark that the relation $\Phi_h(t, y) = \tilde{y}(t+h)$ is nothing other than the Euler-MacLaurin summation formula.

As a particular example we choose the function

$$f(t) = \frac{5}{1 + 25t^2}.$$

Figure 7.1 shows the numerical solution and the exact solution of the modified equation truncated at different values of N . For $h = 0.2$, there is an excellent agreement for $N \leq 12$, whereas oscillations begin to appear from $N = 14$ onwards. For the halved step size $h = 0.1$, the oscillations become visible for N twice as large.

¹ Observe that after adding the equation $\dot{t} = 1$, $t(0) = 0$, we get for $Y = (t, y)^T$ the autonomous differential equation $\dot{Y} = F(Y)$ with $F(Y) = (1, f(t))^T$. Hence, all results of this chapter are applicable.

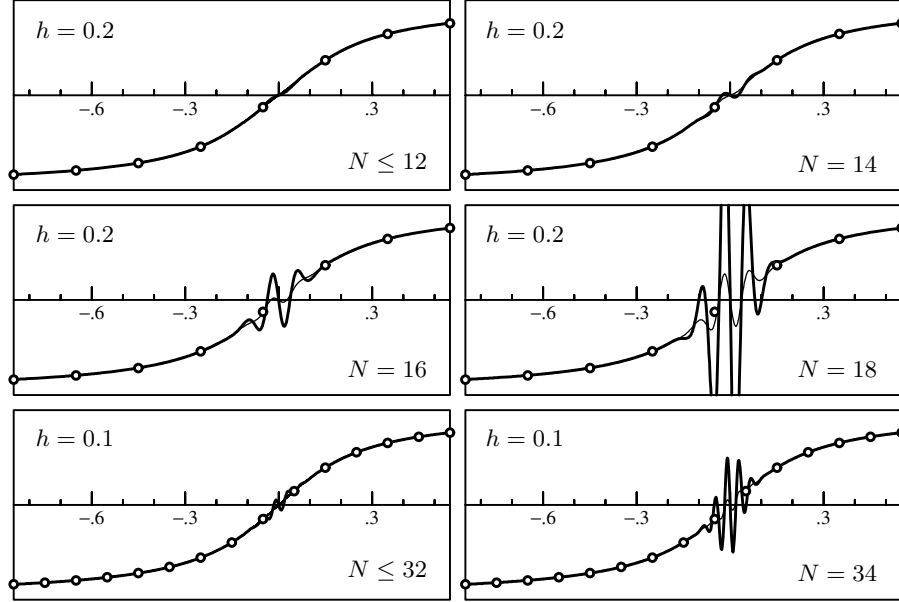


Fig. 7.1. Numerical solution with the trapezoidal rule compared to the solution of the truncated modified equation for $h = 0.2$ (upper four pictures), and for $h = 0.1$ (lower two pictures)

The main ingredient of a rigorous backward error analysis is an analyticity assumption on the differential equation $\dot{y} = f(y)$ and on the method. Throughout this section we assume that $f(y)$ is analytic in a complex neighbourhood of y_0 and that

$$\|f(y)\| \leq M \quad \text{for} \quad \|y - y_0\| \leq 2R \quad (7.2)$$

i.e., for all y of $B_{2R}(y_0) := \{y \in \mathbb{C}^d; \|y - y_0\| \leq 2R\}$. Our strategy is the following: using (7.2) and Cauchy's estimates we derive bounds for the coefficient functions $d_j(y)$ of (1.3) on $B_R(y_0)$ (Sect. IX.7.1), then we estimate the functions $f_j(y)$ of the modified differential equation on $B_{R/2}(y_0)$ (Sect. IX.7.2), and finally we search for a suitable truncation for the formal series (1.1) and we prove the closeness of the numerical solution to the exact solution of the truncated modified equation (Sect. IX.7.3).

IX.7.1 Estimation of the Derivatives of the Numerical Solution

If we apply a numerical method to $\dot{y} = f(y)$ with analytic $f(y)$, the expression $\Phi_h(y)$ will usually be analytic in a neighbourhood of $h = 0$ and $y \in B_R(y_0)$. Consequently, the coefficients $d_j(y)$ of the Taylor series expansion

$$\Phi_h(y) = y + hf(y) + h^2 d_2(y) + h^3 d_3(y) + \dots \quad (7.3)$$

are also analytic and the functions $d_j(y)$ can be estimated by the use of Cauchy's inequalities. Let us demonstrate this for Runge–Kutta methods.

Theorem 7.2. *For a Runge–Kutta method (II.1.4) let*

$$\mu = \sum_{i=1}^s |b_i|, \quad \kappa = \max_{i=1, \dots, s} \sum_{j=1}^s |a_{ij}|. \quad (7.4)$$

If $f(y)$ is analytic in the complex ball $B_{2R}(y_0)$ and satisfies (7.2), then the coefficient functions $d_j(y)$ of (7.3) are analytic in $B_R(y_0)$ and satisfy

$$\|d_j(y)\| \leq \mu M \left(\frac{2\kappa M}{R} \right)^{j-1} \quad \text{for} \quad \|y - y_0\| \leq R. \quad (7.5)$$

Proof. For $y \in B_{3R/2}(y_0)$ and $\|\Delta y\| \leq 1$ the function $\alpha(z) = f(y + z\Delta y)$ is analytic for $|z| \leq R/2$ and bounded by M . Cauchy's estimate therefore yields

$$\|f'(y)\Delta y\| = \|\alpha'(0)\| \leq 2M/R.$$

Consequently, $\|f'(y)\| \leq 2M/R$ for $y \in B_{3R/2}(y_0)$ in the operator norm.

For $y \in B_R(y_0)$, the Runge–Kutta method (II.1.4) requires the solution of the nonlinear system $g_i = y + h \sum_{j=1}^s a_{ij} f(g_j)$, which can be solved by fixed point iteration. If $|h|2\kappa M/R \leq \gamma < 1$, it represents a contraction on the closed set $\{(g_1, \dots, g_s); \|g_i - y\| \leq R/2\}$ and possesses a unique solution. Consequently, the method is analytic for $|h| \leq \gamma R/(2\kappa M)$ and $y \in B_R(y_0)$. This implies that the functions $d_j(y)$ of (7.3) are also analytic. Furthermore, $\|\Phi_h(y) - y\| \leq |h|\mu M$ for $y \in B_R(y_0)$ so that, again by Cauchy's estimate,

$$\|d_j(y)\| = \frac{1}{j!} \left\| \frac{d^j}{dh^j} (\Phi_h(y) - y) \right\|_{h=0} \leq \mu M \left(\frac{2\kappa M}{\gamma R} \right)^{j-1}$$

for $j \geq 1$. The statement is then obtained by considering the limit $\gamma \rightarrow 1$. \square

Due to the consistency condition $\sum_{i=1}^s b_i = 1$, methods with positive weights b_i all satisfy $\mu = 1$. The values μ, κ of some classes of Runge–Kutta methods are given in Table 7.1 (those for the Gauss methods and for the Lobatto IIIA methods have been checked for $s \leq 9$ and $s \leq 5$, respectively).

Estimates of the type (7.5), possibly with a different interpretation of M and R , hold for all one-step methods which are analytic in h and y , e.g., partitioned Runge–Kutta methods, splitting and composition methods, projection methods, Lie group methods,

Table 7.1. The constants μ and κ of formula (7.4)

method	μ	κ	method	μ	κ
explicit Euler	1	0	implicit Euler	1	1
implicit midpoint	1	1/2	trapezoidal rule	1	1
Gauss methods	1	c_s	Lobatto IIIA	1	1

IX.7.2 Estimation of the Coefficients of the Modified Equation

At the beginning of this chapter we gave an explicit formula for the first coefficient functions of the modified differential equation (see (1.4)). Using the Lie derivative

$$(D_i g)(y) = g'(y) f_i(y) \quad (7.6)$$

(cf. (VI.5.2)) and $f_1(y) := f(y)$, these formulas can be written as

$$\begin{aligned} f_2(y) &= d_2(y) - \frac{1}{2!} (D_1 f_1)(y) \\ f_3(y) &= d_3(y) - \frac{1}{3!} (D_1^2 f_1)(y) - \frac{1}{2!} (D_2 f_1 + D_1 f_2)(y). \end{aligned}$$

We have the following recurrence relation for the general case.

Lemma 7.3. *If the numerical method has an expansion of the form (7.3), then the functions $f_j(y)$ of the modified differential equation (1.1) satisfy*

$$f_j(y) = d_j(y) - \sum_{i=2}^j \frac{1}{i!} \sum_{k_1 + \dots + k_i = j} \left(D_{k_1} \dots D_{k_{i-1}} f_{k_i} \right)(y),$$

where $k_m \geq 1$ for all m . Observe that the right-hand expression only involves $f_k(y)$ with $k < j$.

Proof. The solution of the modified equation (1.1) with initial value $y(t) = y$ can be formally written as (cf. (1.2))

$$\tilde{y}(t+h) = y + \sum_{i \geq 1} \frac{h^i}{i!} D^{i-1} F(y),$$

where $F(y) = f_1(y) + h f_2(y) + h^2 f_3(y) + \dots$ stands for the modified equation, and $hD = hD_1 + h^2 D_2 + h^3 D_3 + \dots$ for the corresponding Lie operator. We expand the formal sums and obtain

$$\tilde{y}(t+h) = y + \sum_{i \geq 1} \frac{1}{i!} \sum_{k_1, \dots, k_i} h^{k_1 + \dots + k_i} \left(D_{k_1} \dots D_{k_{i-1}} f_{k_i} \right)(y), \quad (7.7)$$

where all $k_m \geq 1$. Comparing like powers of h in (7.3) and (7.7) yields the desired recurrence relations for the functions $f_j(y)$. \square

To get bounds for $\|f_j(y)\|$, we have to estimate repeatedly expressions like $\|(D_i g)(y)\|$. The following variant of Cauchy's estimate will be extremely useful.

Lemma 7.4. *For analytic functions $f_i(y)$ and $g(y)$ we have for $0 \leq \sigma < \rho$ the estimate*

$$\|D_i g\|_\sigma \leq \frac{1}{\rho - \sigma} \cdot \|f_i\|_\sigma \cdot \|g\|_\rho.$$

Here, $\|g\|_\rho := \max\{\|g(y)\|; y \in B_\rho(y_0)\}$ and $\|f_i\|_\sigma, \|D_i g\|_\sigma$ are defined similarly.

Proof. For a fixed $y \in B_\sigma(y_0)$ the function $\alpha(z) = g(y + zf_i(y))$ is analytic for $\|z\| \leq \varepsilon := (\rho - \sigma)/M$ with $M := \|f_i\|_\sigma$. Since $\alpha'(0) = g'(y)f_i(y) = (D_i g)(y)$, we get from Cauchy's estimate that

$$\|(D_i g)(y)\| = \|\alpha'(0)\| \leq \frac{1}{\varepsilon} \sup_{|z| \leq \varepsilon} \|\alpha(z)\| \leq \frac{M}{\rho - \sigma} \cdot \|g\|_\rho.$$

This proves the statement. \square

We are now able to estimate the coefficients $f_j(y)$ of the modified differential equation.

Theorem 7.5. *Let $f(y)$ be analytic in $B_{2R}(y_0)$, let the Taylor series coefficients of the numerical method (7.3) be analytic in $B_R(y_0)$, and assume that (7.2) and (7.5) are satisfied. Then, we have for the coefficients of the modified differential equation*

$$\|f_j(y)\| \leq \ln 2 \cdot \eta M \left(\frac{\eta M j}{R} \right)^{j-1} \quad \text{for} \quad \|y - y_0\| \leq R/2, \quad (7.8)$$

where $\eta = 2 \max(\kappa, \mu/(2 \ln 2 - 1))$.

Proof. We fix an index, say J , and we estimate (in the notation of Lemma 7.4)

$$\|f_j\|_{R-(j-1)\delta} \quad \text{for} \quad j = 1, 2, \dots, J,$$

where $\delta = R/(2(J-1))$. This will then lead to the desired estimate for $\|f_J\|_{R/2}$.

In the following we abbreviate $\|\cdot\|_{R-(j-1)\delta}$ by $\|\cdot\|_j$. Using repeatedly Cauchy's estimate of Lemma 7.4 we get for $k_1 + \dots + k_i = j$ that

$$\begin{aligned} \|D_{k_1} \dots D_{k_{i-1}} f_{k_i}\|_j &\leq \frac{1}{\delta} \|f_{k_1}\|_j \|D_{k_2} \dots D_{k_{i-1}} f_{k_i}\|_{j-1} \\ &\leq \dots \leq \frac{1}{\delta^{i-1}} \|f_{k_1}\|_j \|f_{k_2}\|_{j-1} \dots \|f_{k_i}\|_{j-i+1} \\ &\leq \frac{1}{\delta^{i-1}} \|f_{k_1}\|_{k_1} \|f_{k_2}\|_{k_2} \dots \|f_{k_i}\|_{k_i}. \end{aligned}$$

The last inequality follows from $\|g\|_j \leq \|g\|_l$ for $l \leq j$, which is an immediate consequence of $B_{R-(j-1)\delta}(y_0) \subset B_{R-(l-1)\delta}(y_0)$. It therefore follows from Lemma 7.3 that

$$\|f_j\|_j \leq \|d_j\|_j + \sum_{i=2}^j \frac{1}{i!} \sum_{k_1+\dots+k_i=j} \frac{1}{\delta^{i-1}} \|f_{k_1}\|_{k_1} \|f_{k_2}\|_{k_2} \dots \|f_{k_i}\|_{k_i}.$$

By induction on j ($1 \leq j \leq J$) we obtain that $\|f_j\|_j \leq \delta \beta_j$, where β_j is defined by

$$\beta_j = \frac{\mu M}{\delta} \left(\frac{2\kappa M}{R} \right)^{j-1} + \sum_{i=2}^j \frac{1}{i!} \sum_{k_1+\dots+k_i=j} \beta_{k_1} \beta_{k_2} \dots \beta_{k_i}. \quad (7.9)$$

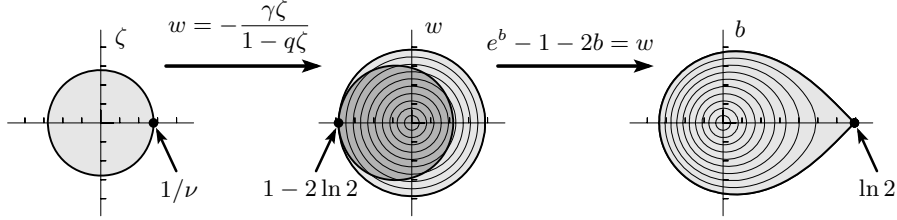


Fig. 7.2. Complex functions of the proof of Theorem 7.5 ($\gamma = q = 1$)

Observe that β_j is defined for all $j \geq 1$. We let $b(\zeta) = \sum_{j \geq 1} \beta_j \zeta^j$ be its generating function and we obtain (by multiplying (7.9) with ζ^j and summing over $j \geq 1$)

$$b(\zeta) = \frac{\gamma\zeta}{1 - q\zeta} + \sum_{j \geq 2} \frac{1}{j!} b(\zeta)^j = \frac{\gamma\zeta}{1 - q\zeta} + e^{b(\zeta)} - 1 - b(\zeta), \quad (7.10)$$

where we have used the abbreviations $\gamma := \mu M / \delta$ and $q := 2\kappa M / R$.

Whenever $e^{b(\zeta)} \neq 2$ (i.e., for $\zeta \neq (2b-1)/(\gamma+q(2b-1))$ with $b = \ln 2 + 2k\pi i$) the implicit function theorem can be applied to (7.10). This implies that $b(\zeta)$ is analytic in a disc with radius $1/\nu = (2 \ln 2 - 1)/(\gamma + q(2 \ln 2 - 1))$ and centre at the origin. On the disc $|\zeta| \leq 1/\nu$, the solution $b(\zeta)$ of (7.10) with $b(0) = 0$ is bounded by $\ln 2$. This is seen as follows (Fig. 7.2): with the function $w = -\gamma\zeta/(1 - q\zeta)$ the disc $|\zeta| \leq 1/\nu$ is mapped into a disc which, for all possible choices of $\gamma \geq 0$ and $q \geq 0$, lies in $|w| \leq 2 \ln 2 - 1$. The image of this disc under the mapping $b(w)$ defined by $e^b - 1 - 2b = w$ and $b(0) = 0$ is completely contained in the disc $|b| \leq \ln 2$. Cauchy's inequalities therefore imply $|\beta_j| \leq \ln 2 \cdot \nu^j$, and we get

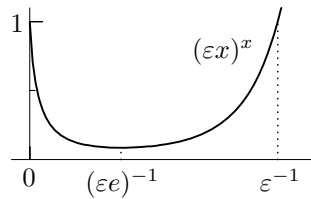
$$\|f_J\|_{R/2} = \|f_J\|_J \leq \delta \beta_J \leq \ln 2 \cdot \delta \cdot \nu^J.$$

Since $\nu = q + \gamma/(2 \ln 2 - 1) \leq \eta M J / R$ with η given by $\eta = 2 \max(\kappa, \mu/(2 \ln 2 - 1))$ and $\delta \nu \leq \eta M$, this proves the statement for J . \square

IX.7.3 Choice of N and the Estimation of the Local Error

To get rigorous estimates, we truncate the modified differential equation (1.1), and we consider

$$\dot{\tilde{y}} = F_N(\tilde{y}), \quad F_N(\tilde{y}) = f(\tilde{y}) + hf_2(\tilde{y}) + \dots + h^{N-1}f_N(\tilde{y}) \quad (7.11)$$



with initial value $\tilde{y}(0) = y_0$. It is common in the theory of asymptotic expansions to truncate the series at the index where the corresponding term is minimal. Motivated by the bound (7.8) and by the fact that $(\varepsilon x)^x$ admits a minimum for $x = (\varepsilon e)^{-1}$ (see the picture to the left with $\varepsilon = 0.15$), we suppose that the truncation index N satisfies

$$hN \leq h_0 \quad \text{with} \quad h_0 = \frac{R}{e\eta M}. \quad (7.12)$$

Under the less restrictive assumption $hN \leq eh_0$, the estimates (7.2) and (7.8) imply for $\|y - y_0\| \leq R/2$ that

$$\begin{aligned} \|F_N(y)\| &\leq M \left(1 + \eta \ln 2 \sum_{j=2}^N \left(\frac{\eta M j h}{R} \right)^{j-1} \right) \\ &\leq M \left(1 + \eta \ln 2 \sum_{j=2}^N \left(\frac{j}{N} \right)^{j-1} \right) \leq M (1 + 1.65 \eta). \end{aligned} \quad (7.13)$$

One can check that the sum in the lower formula of (7.13) is maximal for $N = 7$ and bounded by 2.38. For a p th order method we obtain under the same assumptions

$$\|F_N(y) - f(y)\| \leq c M h^p, \quad (7.14)$$

where c depends only on the method.

Theorem 7.6. *Let $f(y)$ be analytic in $B_{2R}(y_0)$, let the coefficients $d_j(y)$ of the method (7.3) be analytic in $B_R(y_0)$, and assume that (7.2) and (7.5) hold. If $h \leq h_0/4$ with $h_0 = R/(e\eta M)$, then there exists $N = N(h)$ (namely N equal to the largest integer satisfying $hN \leq h_0$) such that the difference between the numerical solution $y_1 = \Phi_h(y_0)$ and the exact solution $\tilde{\varphi}_{N,t}(y_0)$ of the truncated modified equation (7.11) satisfies*

$$\|\Phi_h(y_0) - \tilde{\varphi}_{N,h}(y_0)\| \leq h\gamma M e^{-h_0/h},$$

where $\gamma = e(2 + 1.65\eta + \mu)$ depends only on the method (we have $5 \leq \eta \leq 5.18$ and $\gamma \leq 31.4$ for the methods of Table 7.1).

The quotient $L = M/R$ is an upper bound of the first derivative $f'(y)$ and can be interpreted as a Lipschitz constant for $f(y)$. The condition $h \leq h_0/4$ is therefore equivalent to $hL \leq \text{Const}$, where Const depends only on the method. Because of this condition, Theorem 7.6 requires unreasonably small step sizes for the numerical solution of stiff differential equations.

Proof of Theorem 7.6. We follow here the elegant proof of Benettin & Giorgilli (1994). It is based on the fact that $\Phi_h(y_0)$ (as a convergent series (7.3)) and $\tilde{\varphi}_{N,h}(y_0)$ (as the solution of an analytic differential equation) are both analytic functions of h . Hence,

$$g(h) := \Phi_h(y_0) - \tilde{\varphi}_{N,h}(y_0) \quad (7.15)$$

is analytic in a complex neighbourhood of $h = 0$. By definition of the functions $f_j(y)$ of the modified equation (1.1), the coefficients of the Taylor series for $\Phi_h(y_0)$ and $\tilde{\varphi}_{N,h}(y_0)$ are the same up to the h^N term, but not further due to the truncation of the modified equation. Consequently, the function $g(h)$ contains the factor h^{N+1} ,

and the maximum principle for analytic functions, applied to $g(h)/h^{N+1}$, implies that

$$\|g(h)\| \leq \left(\frac{h}{\varepsilon}\right)^{N+1} \max_{|z| \leq \varepsilon} \|g(z)\| \quad \text{for } 0 \leq h \leq \varepsilon, \quad (7.16)$$

if $g(z)$ is analytic for $|z| \leq \varepsilon$. We shall show that we can take $\varepsilon = eh_0/N$, and we compute an upper bound for $\|g(z)\|$ by estimating separately $\|\Phi_h(y_0) - y_0\|$ and $\|\tilde{\varphi}_{N,h}(y_0) - y_0\|$.

The function $\Phi_z(y_0)$ is given by the series (7.3) which, due to the bounds of Theorem 7.2, converges certainly for $|z| \leq R/(4\kappa M)$, and therefore also for $|z| \leq \varepsilon$ (because $2\kappa \leq \eta$ and $N \geq 4$, which is a consequence of $h_0/h \geq 4$). Hence, it is analytic in $|z| \leq \varepsilon$. Moreover, we have from Theorem 7.2 that $\|\Phi_z(y_0) - y_0\| \leq |z|M(1 + \mu)$ for $|z| \leq \varepsilon$.

Because of the bound (7.13) on $F_N(y)$, which is valid for $y \in B_{R/2}(y_0)$ and for $|h| \leq \varepsilon$, we have $\|\tilde{\varphi}_{N,z}(y_0) - y_0\| \leq |z|M(1 + 1.65\eta)$ as long as the solution $\tilde{\varphi}_{N,z}(y_0)$ stays in the ball $B_{R/2}(y_0)$. Because of $\varepsilon M(1 + 1.65\eta) \leq R/2$, which is a consequence of the definition of ε , of $N \geq 4$, and of $(1 + 1.65\eta) \leq 1.85\eta$ (because for consistent methods $\mu \geq 1$ holds and therefore also $\eta \geq 2/(2 \ln 2 - 1) \geq 5$), this is the case for all $|z| \leq \varepsilon$. In particular, the solution $\tilde{\varphi}_{N,z}(y_0)$ is analytic in $|z| \leq \varepsilon$.

Inserting $\varepsilon = eh_0/N$ and the bound on $\|g(z)\| \leq \|\Phi_z(y_0) - y_0\| + \|\tilde{\varphi}_{N,z}(y_0) - y_0\|$ into (7.16) yields (with $C = 2 + 1.65\eta + \mu$)

$$\|g(h)\| \leq \varepsilon MC \left(\frac{h}{\varepsilon}\right)^{N+1} \leq hMC \left(\frac{h}{\varepsilon}\right)^N = hMC \left(\frac{hN}{eh_0}\right)^N \leq hMCe^{-N},$$

because $hN \leq h_0$. The statement now follows from the fact that $N \leq h_0/h < N + 1$, so that $e^{-N} \leq e \cdot e^{-h_0/h}$. \square

A different approach to a rigorous backward error analysis is developed by Moan (2005). There, the modified differential equation contains an exponentially small time-dependent perturbation, but its flow reproduces the numerical solution without error.

IX.8 Long-Time Energy Conservation

In particular, one easily explains in this way why symplectic algorithms give rise to a good energy conservation, with essentially no accumulation of errors in time. (G. Benettin & A. Giorgilli 1994)

As a first application of Theorem 7.6 we study the long-time energy conservation of symplectic numerical schemes applied to Hamiltonian systems $\dot{y} = J^{-1}\nabla H(y)$. It follows from Theorem 3.1 that the corresponding modified differential equation is also Hamiltonian. After truncation we thus get a modified Hamiltonian

$$\tilde{H}(y) = H(y) + h^p H_{p+1}(y) + \dots + h^{N-1} H_N(y), \quad (8.1)$$

which we assume to be defined on the same open set as the original Hamiltonian H ; see Theorem 3.2 and Sect. IX.4. We also assume that the numerical method satisfies the analyticity bounds (7.5), so that Theorem 7.6 can be applied. The following result is given by Benettin & Giorgilli (1994).

Theorem 8.1. *Consider a Hamiltonian system with analytic $H : D \rightarrow \mathbb{R}$ (where $D \subset \mathbb{R}^{2d}$), and apply a symplectic numerical method $\Phi_h(y)$ with step size h . If the numerical solution stays in the compact set $K \subset D$, then there exist h_0 and $N = N(h)$ (as in Theorem 7.6) such that*

$$\begin{aligned}\tilde{H}(y_n) &= \tilde{H}(y_0) + \mathcal{O}(e^{-h_0/2h}) \\ H(y_n) &= H(y_0) + \mathcal{O}(h^p)\end{aligned}$$

over exponentially long time intervals $nh \leq e^{h_0/2h}$.

Proof. We let $\tilde{\varphi}_{N,t}(y_0)$ be the flow of the truncated modified equation. Since this differential equation is Hamiltonian with \tilde{H} of (8.1), $\tilde{H}(\tilde{\varphi}_{N,t}(y_0)) = \tilde{H}(y_0)$ holds for all times t . From Theorem 7.6 we know that $\|y_{n+1} - \tilde{\varphi}_{N,h}(y_n)\| \leq h\gamma M e^{-h_0/h}$ and, by using a global h -independent Lipschitz constant for \tilde{H} (which exists by Theorem 7.5), we also get $\tilde{H}(y_{n+1}) - \tilde{H}(\tilde{\varphi}_{N,h}(y_n)) = \mathcal{O}(he^{-h_0/h})$. From the identity

$$\tilde{H}(y_n) - \tilde{H}(y_0) = \sum_{j=1}^n \left(\tilde{H}(y_j) - \tilde{H}(y_{j-1}) \right) = \sum_{j=1}^n \left(\tilde{H}(y_j) - \tilde{H}(\tilde{\varphi}_{N,h}(y_{j-1})) \right)$$

we thus get $\tilde{H}(y_n) - \tilde{H}(y_0) = \mathcal{O}(nhe^{-h_0/h})$, and the statement on the long-time conservation of \tilde{H} is an immediate consequence. The statement for the Hamiltonian H follows from (8.1), because $H_{p+1}(y) + hH_{p+2}(y) + \dots + h^{N-p-1}H_N(y)$ is uniformly bounded on K independently of h and N . This follows from the proof of Lemma VI.2.7 and from the estimates of Theorem 7.5. \square

Example 8.2. Let us check explicitly the assumptions of Theorem 8.1 for the pendulum problem $\dot{q} = p$, $\dot{p} = -\sin q$. The vector field $f(p, q) = (p, -\sin q)^T$ is also well-defined for complex p and q , and it is analytic everywhere on \mathbb{C}^2 . We let K be a compact subset of $\{(p, q) \in \mathbb{R}^2 ; |p| \leq c\}$. As a consequence of $|\sin q| \leq e^{|\operatorname{Im} q|}$, we get the bound

$$\|f(p, q)\| \leq \sqrt{c^2 + 4R^2 + e^{2R}}$$

for $\|(p, q) - (p_0, q_0)\| \leq 2R$ and $(p_0, q_0) \in K$. If we choose $c \leq 2$, $R = 1$, and $M = 4$, the value h_0 of Theorem 7.6 is given by $h_0 = 1/4e\eta \approx 0.018$ for the methods of Table 7.1. For step sizes that are smaller than $h_0/20$, Theorem 8.1 guarantees that the numerical Hamiltonian is well conserved on intervals $[0, T]$ with $T \approx e^{10} \approx 2 \cdot 10^4$.

The numerical experiment of Fig. 8.1 shows that the estimates for h_0 are often too pessimistic. We have drawn 200 000 steps of the numerical solution of the

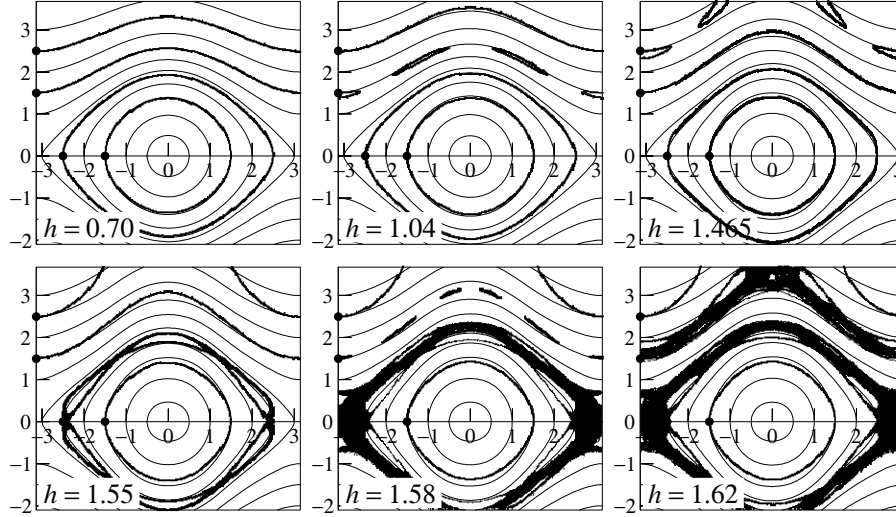


Fig. 8.1. Numerical solutions of the implicit midpoint rule with large step sizes

implicit midpoint rule for various step sizes h and for initial values $(p_0, q_0) = (0, -1.5)$, $(p_0, q_0) = (0, -2.5)$, $(p_0, q_0) = (1.5, -\pi)$, and $(p_0, q_0) = (2.5, -\pi)$. They are compared to the contour lines of the truncated modified Hamiltonian

$$\tilde{H}(p, q) = \frac{p^2}{2} - \cos q + \frac{h^2}{48} (\cos(2q) - 2p^2 \cos q).$$

This shows that for step sizes as large as $h \leq 0.7$ the Hamiltonian \tilde{H} is extremely well conserved. Beyond this value, the dynamics of the numerical method soon turns into chaotic behaviour (see also Yoshida (1993) and Hairer, Nørsett & Wanner (1993), page 336).

Theorem 8.1 explains the near conservation of the Hamiltonian with the symplectic Euler method, the implicit midpoint rule and the Störmer–Verlet method as observed in the numerical experiments of Chap. I: in Fig. I.1.4 for the pendulum problem, in Fig. I.2.3 for the Kepler problem, and in Fig. I.4.1 for the frozen argon crystal.

The linear drift of the numerical Hamiltonian for non-symplectic methods can be explained by a computation similar to that of the proof of Theorem 8.1. From a Lipschitz condition of the Hamiltonian and from the standard local error estimate, we obtain $H(y_{n+1}) - H(\varphi_h(y_n)) = \mathcal{O}(h^{p+1})$. Since $H(\varphi_h(y_n)) = H(y_n)$, a summation of these terms leads to

$$H(y_n) - H(y_0) = \mathcal{O}(th^p) \quad \text{for } t = nh. \quad (8.2)$$

This explains the linear growth in the error of the Hamiltonian observed in Fig. I.2.3 and in Fig. I.4.1 for the explicit Euler method.

IX.9 Modified Equation in Terms of Trees

By Theorem III.1.4 the numerical solution $y_1 = \Phi_h(y_0)$ of a Runge–Kutta method can be written as a B-series

$$\begin{aligned} \Phi_h(y) = & y + hf(y) + h^2 a(\bullet)(f'f)(y) \\ & + h^3 \left(\frac{1}{2} a(\vee) f''(f, f)(y) + a(\curvearrowright) f'f'f(y) \right) + \dots \end{aligned} \quad (9.1)$$

For consistent methods, i.e., methods of order at least 1, we always have $a(\bullet) = 1$, so that the coefficient of h is equal to $f(y)$. In this section we exploit this special structure of $\Phi_h(y)$ in order to get practical formulas for the coefficient functions of the modified differential equation. Using (9.1) instead of (1.3), the equations (1.4) yield

$$\begin{aligned} f_2(y) &= \left(a(\bullet) - \frac{1}{2} \right) (f'f)(y) \\ f_3(y) &= \frac{1}{2} \left(a(\vee) - a(\bullet) + \frac{1}{6} \right) f''(f, f)(y) \\ &\quad + \left(a(\curvearrowright) - a(\bullet) + \frac{1}{3} \right) f'f'f(y). \end{aligned} \quad (9.2)$$

Continuing this computation, one is quickly convinced of the general formula

$$f_j(y) = \sum_{|\tau|=j} \frac{b(\tau)}{\sigma(\tau)} F(\tau)(y), \quad (9.3)$$

so that the modified equation (1.1) becomes

$$\dot{\tilde{y}} = \sum_{\tau \in T} \frac{h^{|\tau|-1}}{\sigma(\tau)} b(\tau) F(\tau)(\tilde{y}) \quad (9.4)$$

with $b(\bullet) = 1$, $b(\bullet) = a(\bullet) - \frac{1}{2}$, etc. Since the coefficients $\sigma(\tau)$ are known from Definition III.1.7, all we have to do is to find suitable recursion formulas for the real coefficients $b(\tau)$.

IX.9.1 B-Series of the Modified Equation

Recurrence formulas for the coefficients $b(\tau)$ in (9.4) were first given by Hairer (1994) and by Calvo, Murua & Sanz-Serna (1994). We follow here the approach of Hairer (1999), which uses the Lie-derivative of B-series and thus simplifies the construction of the coefficients.

We make use of the notion of ordered trees introduced in Sect. III.1.3. For a given tree τ we define the set of all *splittings* as

$$SP(\tau) = \{ \theta \in OST(\tau) ; \tau \setminus \theta \text{ consists of only one element} \}. \quad (9.5)$$

Here, $OST(\tau) = OST(\omega(\tau))$ is the set of ordered subtrees as defined in (III.1.33).

Lemma 9.1 (Lie-Derivative of B-series). *Let $b(\tau)$ (with $b(\emptyset) = 0$) and $c(\tau)$ be the coefficients of two B-series, and let $y(t)$ be a formal solution of the differential equation $h\dot{y}(t) = B(b, y(t))$. The Lie derivative of the function $B(c, y)$ with respect to the vector field $B(b, y)$ is again a B-series*

$$h \frac{d}{dt} B(c, y(t)) = B(\partial_b c, y(t)).$$

Its coefficients are given by $\partial_b c(\emptyset) = 0$ and for $|\tau| \geq 1$ by

$$\partial_b c(\tau) = \sum_{\theta \in SP(\tau)} c(\theta) b(\tau \setminus \theta). \quad (9.6)$$

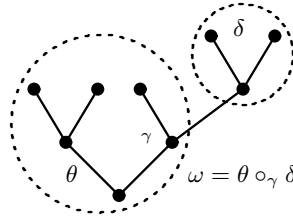


Fig. 9.1. Splitting of an ordered tree ω into a subtree θ and $\{\delta\} = \omega \setminus \theta$

Proof. For the proof of this lemma it is convenient to work with ordered trees $\omega \in OT$. Since $\nu(\tau)$ of (III.1.31) denotes the number of possible orderings of a tree $\tau \in T$, a sum $\sum_{\tau \in T} \cdot / \cdot$ becomes $\sum_{\omega \in OT} \nu(\omega)^{-1} \cdot / \cdot$.

For the computation of the Lie derivative of $B(c, y)$ we have to differentiate the elementary differential $F(\theta)(y(t))$ with respect to t . Using Leibniz' rule, this yields $|\theta|$ terms, one for every vertex of θ . Then we insert the series $B(b, y(t))$ for $h\dot{y}(t)$. This means that all the trees δ appearing in $B(b, y(t))$ are attached with a new branch to the distinguished vertex. Written out as formulas, this gives

$$h \frac{d}{dt} B(c, y(t)) = \sum_{\theta \in OT \cup \{\emptyset\}} \frac{h^{|\theta|} c(\theta)}{\nu(\theta) \sigma(\theta)} \sum_{\gamma} \sum_{\delta \in OT} \frac{h^{|\delta|} b(\delta)}{\nu(\delta) \sigma(\delta)} F(\theta \circ_{\gamma} \delta)(y(t)),$$

where \sum_{γ} is a sum over all vertices of θ , and $\theta \circ_{\gamma} \delta$ is the ordered tree obtained when attaching the root of δ with a new branch to γ (see Fig. 9.1). We choose one of the $n(\gamma) + 1$ possibilities of doing this, where $n(\gamma)$ denotes the number of upwards leaving branches of θ at the vertex γ . We now collect the terms with equal ordered tree $\omega = \theta \circ_{\gamma} \delta$, and notice that $\nu(\theta) \sigma(\theta) = \kappa(\theta)$ with $\kappa(\theta)$ given by (III.1.32). This gives

$$h \frac{d}{dt} B(c, y(t)) = \sum_{\omega \in OT} h^{|\omega|} \left(\sum_{\theta \circ_{\gamma} \delta = \omega} \frac{c(\theta) b(\delta)}{(n(\gamma) + 1) \kappa(\theta) \kappa(\delta)} \right) F(\omega)(y(t)),$$

where $\sum_{\theta \circ_{\gamma} \delta = \omega}$ is over all triplets (θ, γ, δ) such that $\theta \circ_{\gamma} \delta = \omega$. Because of $\kappa(\omega) = \kappa(\theta) \kappa(\delta) (n(\gamma) + 1)$, we obtain

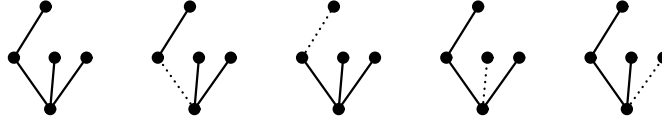


Fig. 9.2. Illustration of the formula (9.6) for an ordered tree with 5 vertices

$$\begin{aligned}
 h \frac{d}{dt} B(c, y(t)) &= \sum_{\omega \in OT} \frac{h^{|\omega|}}{\kappa(\omega)} \left(\sum_{\theta \circ_{\gamma} \delta = \omega} c(\theta) b(\delta) \right) F(\omega)(y(t)) \\
 &= \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} \left(\sum_{\theta \in SP(\tau)} c(\theta) b(\tau \setminus \theta) \right) F(\tau)(y(t)),
 \end{aligned}$$

which proves the statement. \square

Let us illustrate this proof and the formula (9.6) with an ordered tree having 5 vertices. All possible splittings $\omega = \theta \circ_{\gamma} \delta$ are given in Fig.9.2. Notice that θ may be the empty tree \emptyset , and that always $|\delta| \geq 1$. We see that the tree ω is obtained in several ways: (i) differentiation of $F(\emptyset)(y) = y$ and adding $F(\omega)(y)$ as argument, (ii) differentiation of the factor corresponding to the root in $F(\theta)(y) = f''(f, f)(y)$ and adding $F(\text{root})(y) = (f'f)(y)$, (iii) differentiation of all f 's in $F(\theta)(y) = f'''(f, f, f)(y)$ and adding $F(\bullet)(y) = f(y)$, and finally, (iv) differentiation of the factor for the root in $F(\theta)(y) = f''(f'f, f)(y)$ and adding $F(\bullet)(y) = f(y)$. This proves that

$$\partial_b c(\text{root}) = c(\emptyset) b(\text{root}) + c(\text{root}) b(\text{root}) + c(\text{root}) b(\bullet) + 2 c(\text{root}) b(\bullet).$$

For the trees up to order 3 the formulas for $\partial_b c$ are:

$$\begin{aligned}
 \partial_b c(\bullet) &= c(\emptyset) b(\bullet) \\
 \partial_b c(\text{root}) &= c(\emptyset) b(\text{root}) + c(\bullet) b(\bullet) \\
 \partial_b c(\text{root}) &= c(\emptyset) b(\text{root}) + 2 c(\text{root}) b(\bullet) \\
 \partial_b c(\text{root}) &= c(\emptyset) b(\text{root}) + c(\bullet) b(\text{root}) + c(\text{root}) b(\bullet).
 \end{aligned}$$

The above lemma permits us to get recursion formulas for the coefficients $b(\tau)$ of the modified differential equation (9.4).

Theorem 9.2. *If the method $\Phi_h(y)$ is given by (9.1), the functions $f_j(y)$ of the modified differential equation (1.1) satisfy (9.3), where the real coefficients $b(\tau)$ are recursively defined by $b(\emptyset) = 0$, $b(\bullet) = 1$ and*

$$b(\tau) = a(\tau) - \sum_{j=2}^{|\tau|} \frac{1}{j!} \partial_b^{j-1} b(\tau). \quad (9.7)$$

Here, ∂_b^{j-1} is the $(j-1)$ -th iterate of the Lie-derivative ∂_b defined in Lemma 9.1.

Proof. The right-hand side of the modified equation (9.4) is the B-series $B(b, \tilde{y}(t))$ divided by h . It therefore follows from an iterative application of Lemma 9.1 that

$$h^j \tilde{y}^{(j)}(t) = B(\partial_b^{j-1} b, \tilde{y}(t)),$$

so that by Taylor series expansion $\tilde{y}(t+h) = y + B(\sum_{j \geq 1} \frac{1}{j!} \partial_b^{j-1} b, y)$, where $y := \tilde{y}(t)$. Since we have to determine the coefficients $b(\tau)$ in such a way that $\tilde{y}(t+h) = \Phi_h(y) = B(a, y)$, a comparison of the two B-series gives $\sum_{j \geq 1} \frac{1}{j!} \partial_b^{j-1} b(\tau) = a(\tau)$. This proves the statement, because $\partial_b^0 b(\tau) = b(\tau)$ for $\tau \in T$, and $\partial_b^{j-1} b(\tau) = 0$ for $j > |\tau|$ (as a consequence of $b(\emptyset) = 0$). \square

We present in Table 9.1 the formula (9.7) for trees up to order 3.

Table 9.1. Examples of formula (9.7)

$\tau = \bullet$	$b(\bullet) = a(\bullet)$
$\tau = \text{J}$	$b(\text{J}) = a(\text{J}) - \frac{1}{2} b(\bullet)^2$
$\tau = \text{V}$	$b(\text{V}) = a(\text{V}) - b(\text{J})b(\bullet) - \frac{1}{3} b(\bullet)^3$
$\tau = \text{J}'$	$b(\text{J}') = a(\text{J}') - b(\text{J})b(\bullet) - \frac{1}{6} b(\bullet)^3$

We next consider the case when a symplectic method is applied to a Hamiltonian system $\dot{y} = J^{-1} \nabla H(y)$. It follows from Theorem 3.1 that the modified equation is again Hamiltonian. What does this imply for the coefficients of (9.4)?

Theorem 9.3. *Suppose that for all Hamiltonians $H(y)$ the modified vector field (9.4), truncated after an arbitrary power of h , is (locally) Hamiltonian. Then,*

$$b(u \circ v) + b(v \circ u) = 0 \quad \text{for all } u, v \in T. \quad (9.8)$$

Proof. Let $\tilde{\varphi}_{N,t}(y_0)$ be the flow of the modified differential equation (9.4), truncated after the h^{N-1} terms. It is symplectic for all t , and in particular for $t = h$. As a consequence of the proof of Theorem 9.2 we obtain that $\tilde{\varphi}_{N,h}(y_0)$ is a symplectic B-series $B(a_N, y_0)$. The coefficients $a_N(\tau)$ are given by (9.7), where $b(\tau)$ is replaced with 0 for $|\tau| > N$. For $u, v \in T$ with $|u| + |v| = N$ we therefore have

$$b(u \circ v) = a_N(u \circ v) - a_{N-1}(u \circ v).$$

Since $a_N(\tau) = a_{N-1}(\tau)$ for $|\tau| < N$, formula (9.8) is an immediate consequence of Theorem VI.7.6. \square

Remark 9.4. Let $G = \{a : T \rightarrow \mathbb{R} \mid a(\emptyset) = 1\}$ be the Butcher group (see Sect. III.1.5), and consider the mapping $S : G \rightarrow \mathbb{R}$ defined by

$$S(a) = a(u \circ v) + a(v \circ u) - a(u) \cdot a(v).$$

If we denote by $e \in G$ the element corresponding to the identity (i.e., $e(\emptyset) = 1$ and $e(\tau) = 0$ for $|\tau| \geq 1$), we have for its derivative

$$S'(e)b = b(u \circ v) + b(v \circ u).$$

Hence, coefficient mappings $b(\tau)$ satisfying (9.8) lie in the tangent space at $e(\tau)$ of the symplectic subgroup of G (i.e., $a \in G$ satisfying (VI.7.4)). This is in complete analogy to the fact that Hamiltonian vector fields can be considered as elements of the tangent space at the identity of the group of symplectic diffeomorphisms (see also Exercises 15 and 16).

IX.9.2 Elementary Hamiltonians

If the modified differential equation (9.4) is Hamiltonian, can we find explicit formulas for $\tilde{H}(y)$? Let us start with an easy example, the implicit midpoint rule. Written as a B-series (9.1), its coefficients are $a(\tau) = 2^{1-|\tau|}$ (cf. Exercise 8) so that the first coefficient functions (9.2) of the modified equation satisfy $f_2(y) = 0$ and

$$f_3(y) = \frac{1}{24} \left(2(f' f' f)(y) - f''(f, f)(y) \right). \quad (9.9)$$

Since $f(y) = J^{-1} \nabla H(y)$, differentiation of

$$H_3(y) = -\frac{1}{24} H''(y) \left(J^{-1} \nabla H(y), J^{-1} \nabla H(y) \right) \quad (9.10)$$

shows that $f_3(y) = J^{-1} \nabla H_3(y)$, and we have found an explicit expression of the Hamiltonian corresponding to the vector field $f_3(y)$. It is recommended to compute also $f_5(y)$ and to try to find $H_5(y)$ such that $f_5(y) = J^{-1} \nabla H_5(y)$. Such computations lead to expressions that have been introduced in a different context by Sanz-Serna & Abia (1991). They call them *canonical elementary differentials*.

Definition 9.5 (Elementary Hamiltonians). For a given smooth function $H : D \rightarrow \mathbb{R}$ (with open $D \subset \mathbb{R}^{2d}$) and for $\tau \in T$ we define the *elementary Hamiltonian* $H(\tau) : D \rightarrow \mathbb{R}$ by

$$H(\bullet)(y) = H(y), \quad H(\tau)(y) = H^{(m)}(y) (F(\tau_1)(y), \dots, F(\tau_m)(y)) \quad (9.11)$$

for $\tau = [\tau_1, \dots, \tau_m]$. Here, $F(\tau_i)(y)$ are elementary differentials corresponding to $f(y) = J^{-1} \nabla H(y)$.

The expression in (9.10) is nothing else than the elementary Hamiltonian corresponding to the tree \mathbf{V} . Our aim is to prove that, for symplectic methods applied to Hamiltonian systems, the coefficient functions (9.3) of the modified differential equation satisfy $f_j(y) = J^{-1} \nabla H_j(y)$, where $H_j(y)$ is a linear combination of elementary Hamiltonians.

Lemma 9.6. *Elementary Hamiltonians satisfy*

$$H(u \circ v)(y) + H(v \circ u)(y) = 0 \quad \text{for all } u, v \in T. \quad (9.12)$$

In particular, we have $H(u \circ u)(y) = 0$ for all $u \in T$.

Proof. This follows immediately from the fact that for $u = [u_1, \dots, u_m] \in T$ and for $v \in T$ we have $H(u \circ v) = H^{(m+1)}(F(u_1), \dots, F(u_m), F(v)) = F(v)^T (\nabla H)^{(m)}(F(u_1), \dots, F(u_m)) = F(v)^T J F(u)$, and from the skew-symmetry of J . \square

The trees $u \circ v$ and $v \circ u$ have the same graph and differ only in the position of the root. The relation (9.12) thus motivates the consideration of the (smallest) equivalence relation on T satisfying

$$u \circ v \sim v \circ u. \quad (9.13)$$

We want to select from each equivalence class, not containing a tree of the form $u \circ u$, exactly one element. This can be done as follows (cf. Chartier, Faou & Murua 2005): we choose a total ordering on the set T that respects the number of vertices, i.e., $u < v$ whenever $|u| < |v|$, and we define

$$\begin{aligned} T^* &= \{\bullet\} \cup \{\tau \mid \tau \text{ cannot be written as } \tau = u \circ v \text{ with } u \leq v\} \\ &= \left\{ \bullet, \begin{array}{c} \diagup \diagdown \\ \bullet \end{array}, \begin{array}{c} \diagup \diagdown \\ \diagup \diagdown \\ \bullet \end{array}, \begin{array}{c} \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \bullet \end{array}, \begin{array}{c} \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \bullet \end{array}, \begin{array}{c} \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \bullet \end{array}, \begin{array}{c} \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \bullet \end{array}, \begin{array}{c} \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \diagup \diagdown \\ \bullet \end{array}, \dots \right\} \end{aligned} \quad (9.14)$$

(for the second line we assume $[\bullet, \bullet] < [[\bullet]]$). Every tree $\tau \in T$ is either equivalent to some $u \circ u$ or to a tree in T^* . This is a consequence of the fact that as long as $\tau = u \circ v$ with $u < v$, it can be changed to $v \circ u$ (what happens only a finite number of times). Moreover, two trees of T^* can never be equivalent.

Lemma 9.7. *For a tree $\tau \in T^*$ we have*

$$J^{-1} \nabla H(\tau)(y) = \sigma(\tau) \sum_{\theta \sim \tau} \frac{(-1)^{\kappa(\tau, \theta)}}{\sigma(\theta)} F(\theta)(y), \quad (9.15)$$

where $\kappa(\tau, \theta)$ is the number of root changes that are necessary to obtain θ from τ .

Proof. We compute $J^{-1} \nabla H(\tau)(y)$. The expression $H(\tau)(y)$ consists of $|\tau|$ factors corresponding to the vertices of τ , each of which has to be differentiated by Leibniz' rule. Differentiation of $H^{(m)}(y)$ (cf. Definition 9.5) and pre-multiplication by the matrix J^{-1} yields $F(\tau)(y)$. Before differentiating the other factors, we bring the corresponding vertex down to the root. In view of Lemma 9.6 this only multiplies $H(\tau)(y)$ by $(-1)^{\kappa(\tau, \theta)}$, and shows that a differentiation of the corresponding factor yields $F(\theta)(y)$. Since $\tau \in T^*$, the number of possibilities to obtain θ from τ by exchanging roots is equal to $\sigma(\tau)/\sigma(\theta)$. This factor has to be included. \square

IX.9.3 Modified Hamiltonian

We are now in the position to give an explicit formula for the Hamiltonian of the modified differential equation provided that the numerical method can be written as a B-series. An extension to partitioned methods will be given in Sect. IX.10.

Theorem 9.8. *Consider a numerical method that can be written as a B-series (9.1), and that is symplectic for every Hamiltonian system $\dot{y} = J^{-1}\nabla H(y)$. Its modified differential equation is then Hamiltonian with*

$$\tilde{H}(y) = H_1(y) + h H_2(y) + h^2 H_3(y) + \dots,$$

where

$$H_j(y) = \sum_{\tau \in T^*, |\tau|=j} \frac{b(\tau)}{\sigma(\tau)} H(\tau)(y), \quad (9.16)$$

and the coefficients $b(\tau)$ are those of Theorem 9.2. Notice that the sum in (9.16) is only over trees in T^* as defined in (9.14).

Proof. We apply the method (9.1) to the Hamiltonian system, so that by Theorem 3.1 the modified differential equation is (locally) Hamiltonian. It therefore follows from Theorem 9.3 that the coefficients $b(\tau)$ of (9.4) satisfy (9.8). This relation implies $b(\theta) = (-1)^{\kappa(\tau, \theta)} b(\tau)$ whenever $\theta \sim \tau$. Inserted into (9.3), an application of Lemma 9.7 proves the statement. \square

Remark 9.9. This theorem gives an explicit formula for the modified Hamiltonian (for methods expressed as B-series). Since the elementary Hamiltonians $H(\tau)(y)$ depend only on derivatives of $H(y)$, this modified Hamiltonian is *globally* defined. For Runge–Kutta methods this provides an alternative approach to the statement of Theorem 3.2.

For the sake of completeness we give in the following theorem a characterization of Hamiltonian vector fields of the form (9.4).

Theorem 9.10. *The differential equation $h\dot{y} = B(b, y)$ with $b(\emptyset) = 0$ is Hamiltonian for all vector fields $f(y) = J^{-1}\nabla H(y)$, if and only if*

$$b(u \circ v) + b(v \circ u) = 0 \quad \text{for all } u, v \in T. \quad (9.17)$$

Proof. The “only if” part follows from Theorem 9.3. The “if” part is a consequence of the proof of Theorem 9.8. \square

IX.9.4 First Integrals Close to the Hamiltonian

We have seen in Sect. IX.9.3 that for symplectic methods the modified differential equation (9.4) based on $f(y) = J^{-1}\nabla H(y)$ is Hamiltonian with a function of the form

$$H(c, y) = \sum_{\tau \in T^*} \frac{h^{|\tau|-1}}{\sigma(\tau)} c(\tau) H(\tau)(y) \quad (9.18)$$

and coefficients $c(\tau) = b(\tau)$. In this section we study whether for non-symplectic methods a function of the form (9.18) can be a first integral of (9.4). This question has been addressed by Faou, Hairer & Pham (2004), and we closely follow their presentation.

Lemma 9.11. *Let $y(t)$ be a solution of the differential equation (9.4) which can be written as $h\dot{y}(t) = B(b, y(t))$. We then have*

$$\frac{d}{dt} H(c, y(t)) = H(\delta_b c, y(t))$$

where $\delta_b c(\bullet) = 0$ and, for $\tau \in T^*$ with $|\tau| > 1$,

$$\delta_b c(\tau) = \sum_{\theta \sim \tau} (-1)^{\kappa(\tau, \theta)} \frac{\sigma(\tau)}{\sigma(\theta)} \sum_{\omega \in T^* \cap SP(\theta)} c(\omega) b(\theta \setminus \omega). \quad (9.19)$$

The first sum is over all trees θ that are equivalent to τ (see (9.13)), and the second sum is over all splittings of θ as in Lemma 9.1 (see Table 9.2).

Proof. The proof is nearly the same as that of Lemma 9.1. The first sum in (9.19) appears, because $H(\theta)(y) = H(\tau)(y)$ for $\theta \sim \tau$ and because the sum in (9.18) is only over trees in T^* . \square

Table 9.2. Formulas for $\delta_b c(\tau)$ for trees $\tau \in T^*$ up to order 6

$$\begin{aligned} \delta_b c(\text{V}) &= -2 c(\bullet) b(\text{I}) \\ \delta_b c(\text{V} \setminus \text{V}) &= 3 c(\text{V}) b(\bullet) - 3 c(\bullet) b(\text{V}) \\ \delta_b c(\text{V} \setminus \text{V} \setminus \text{V}) &= 4 c(\text{V} \setminus \text{V}) b(\bullet) - 4 c(\bullet) b(\text{V} \setminus \text{V}) \\ \delta_b c(\text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V}) &= c(\text{V} \setminus \text{V}) b(\bullet) + c(\text{V}) b(\text{I}) + c(\bullet) b(\text{V} \setminus \text{V}) - 2 c(\bullet) b(\text{V} \setminus \text{V}) \\ \delta_b c(\text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V}) &= 2 c(\bullet) b(\text{V} \setminus \text{V}) - 2 c(\text{V}) b(\text{I}) \\ \delta_b c(\text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V}) &= 5 c(\text{V} \setminus \text{V}) b(\bullet) - 5 c(\bullet) b(\text{V} \setminus \text{V}) \\ \delta_b c(\text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V}) &= 3 c(\text{V} \setminus \text{V}) b(\bullet) + c(\text{V} \setminus \text{V} \setminus \text{V}) b(\bullet) + c(\text{V} \setminus \text{V}) b(\text{I}) \\ &\quad - 3 c(\bullet) b(\text{V} \setminus \text{V}) + c(\bullet) b(\text{V} \setminus \text{V} \setminus \text{V}) \\ \delta_b c(\text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V}) &= 2 c(\text{V} \setminus \text{V}) b(\bullet) + c(\text{V} \setminus \text{V} \setminus \text{V}) b(\bullet) - c(\bullet) b(\text{V} \setminus \text{V} \setminus \text{V}) + 2 c(\bullet) b(\text{V} \setminus \text{V}) \\ \delta_b c(\text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V} \setminus \text{V}) &= 2 c(\text{V} \setminus \text{V}) b(\bullet) - c(\text{V} \setminus \text{V}) b(\bullet) - c(\text{V} \setminus \text{V}) b(\text{I}) - c(\text{V}) b(\text{V}) \\ &\quad - c(\text{V}) b(\text{I}) + 2 c(\bullet) b(\text{V} \setminus \text{V}) + c(\bullet) b(\text{V} \setminus \text{V}) \end{aligned}$$

Corollary 9.12. *The function $H(c, y)$ of (9.18) is a first integral of the differential equation (9.4) for every $H(y)$ if and only if*

$$\delta_b c(\tau) = 0 \quad \text{for all } \tau \in T^*. \quad (9.20)$$

Proof. The sufficiency follows from Lemma 9.11 and the necessity is a consequence of the independence of the elementary Hamiltonians. To prove their independence we have to show that the series (9.18) vanishes for all smooth $H(y)$ only if $c(\tau) = 0$ for all $\tau \in T^*$. With the techniques of the proof of Theorem VI.7.4 one can show that for every tree $\tau \in T^*$ there exists a polynomial Hamiltonian such that the first component of $F(\tau)(0)$ vanishes for all trees except for τ . Differentiating (9.18) and employing Lemma 9.7 proves that $c(\tau) = 0$. \square

Solving the System (9.20). We consider a consistent method, i.e., $b(\bullet) = 1$, and we search for a first integral $H(c, y)$ close to the Hamiltonian, i.e., $c(\bullet) = 1$.

$|\tau| = 3$: The condition (9.20) for $\tau = \mathbf{V}$ implies $b(\mathbf{J}) = 0$, which means that the method has to be of order two.

$|\tau| = 4$: There is only one tree in T^* with four vertices. The corresponding condition can be satisfied by putting $c(\mathbf{V}) = b(\mathbf{V})$.

$|\tau| = 5$: The third condition yields $b(\llbracket \bullet \rrbracket) = 0$. Letting $c(\mathbf{V})$ be such that one of the other two conditions holds, we still have to satisfy

$$b(\mathbf{V}) + b(\mathbf{V}) - 2b(\mathbf{V}) = 0. \quad (9.21)$$

This condition is satisfied for symplectic methods, for which $b(u \circ v) + b(v \circ u) = 0$, and also for symmetric methods, for which $b(\tau) = 0$ for trees with an even order.

$|\tau| = 6$: There are four conditions for three $c(\tau)$ coefficients. Assuming (9.20) for trees with less than five vertices, these four conditions admit a solution if and only if

$$\begin{aligned} 5b(\mathbf{V}) + 5b(\mathbf{V}) + 6b(\mathbf{V}) + 6b(\mathbf{V}) - 12b(\mathbf{V}) + 3b(\mathbf{V}) \\ - 15b(\mathbf{V}) - 3b(\mathbf{V})(b(\mathbf{V}) + b(\mathbf{J})) = 0. \end{aligned} \quad (9.22)$$








This relation is obviously satisfied by every symplectic method. However, as we shall see soon, there are symmetric methods that do not satisfy (9.22).

For various symmetric methods of order 4 (i.e., $b(\tau) = 0$ for $1 < |\tau| < 5$) we compute the coefficients $b(\tau)$ of the leading perturbation term in (9.4) and also the expression (9.22), see Table 9.3. None of the considered methods is symplectic.

Surprisingly, the 3-stage collocation method Lobatto IIIA (see Table II.1.2 for the coefficients) satisfies the condition (9.22). This implies for every Hamiltonian system (reversible or not reversible) that the dominating error term in the numerical Hamiltonian does not have any drift.

The 3-stage Lobatto IIIB method (see Table II.1.4) does not satisfy the condition (9.22). We therefore expect a drift in the numerical Hamiltonian.

Table 9.3. Coefficients $b(\tau)$ and expression (9.22) for methods of order 4

method								(9.22)
Lobatto IIIA	$\frac{1}{120}$	$\frac{1}{240}$	$\frac{1}{480}$	$-\frac{1}{120}$	$-\frac{1}{240}$	$\frac{1}{720}$	$-\frac{1}{360}$	0
Lobatto IIIB	$\frac{1}{120}$	$-\frac{1}{360}$	$-\frac{1}{720}$	$-\frac{1}{120}$	$\frac{1}{360}$	$\frac{1}{720}$	$\frac{1}{240}$	$\frac{1}{48}$

Lemma 9.13. For given $b(\tau), \tau \in T$ satisfying $b(\emptyset) = 0, b(\bullet) = 1$, and for fixed $c(\bullet)$, the linear system (9.20) for $c(\tau), \tau \in T^*$ has at most one solution.

Proof. We prove by induction on $\tau \in T^*$ that $c(\tau)$ is uniquely determined by (9.20). For this we assume that the ordering on T is such that, within trees of the same order, it is increasing when the number of vertices connected to the root decreases, cf. (9.14).

Let $\tau = [\tau_1, \dots, \tau_m, \bullet, \dots, \bullet] \in T^* \setminus \{\bullet\}$ with $|\tau_j| > 1$, and denote by k the number of \bullet 's in this representation. Since the tree $\tau \circ \bullet$ is again in the set T^* , condition (9.20) yields

$$0 = \delta_b c(\tau \circ \bullet) = (k+1)c(\tau)b(\bullet) - (k+1)c(\bullet)b(\tau) + \dots \quad (9.23)$$

For $m = 0$, no further terms are present and $c(\tau)$ is uniquely determined by this relation. For $m > 0$, the three dots in (9.23) represent a linear combination of $c(\mu)b(\nu)$ with $|\mu| < |\tau|$ (which, by the induction hypothesis, are already known) and of $c(\sigma)b(\bullet)$, where $\sigma \in T^*$ is the representant in T^* of the equivalence class for τ' . We use the notation τ' for some tree which is obtained from τ by removing one of the end vertices of τ_j and by adding it to the root of τ .

In general we will have $\tau' \in T^*$ (so that $\sigma = \tau'$), and in this case its number of end vertices connected to the root is larger than that for τ . Hence, $\sigma < \tau$, and the coefficient $c(\sigma)$ is known by the induction hypothesis.

If $\tau' \notin T^*$, what is only possible if $\tau = u \circ v$ with $|u| = |v|$ and $u > v$, we have $\tau' = u' \circ v$ and $u' < v$ (notice that $u' = v$ is not permitted for trees in T^*). In this case we have $\sigma = v \circ u' \in T^*$. Consequently, $c(\tau) = c(u \circ v)$ is expressed in terms of $c(v \circ u')$ and known quantities. Applying the same reasoning to $v \circ u'$ and observing that because of $u > v$ the tree v has at least as many end vertices connected to the root as the tree u , we see that $c(v \circ u')$ is expressed in terms of already determined quantities. \square

The expression (9.20) is bilinear in b and c . Assuming that $h\dot{y} = B(b, y)$ is Hamiltonian, the mapping b has the same degree of freedom as c . It is therefore not astonishing to have the following dual variant of Lemma 9.13.

Lemma 9.14. Let $c(\tau), \tau \in T^*$ be given and assume $c(\bullet) = 1$ and $b(\emptyset) = 0$. Then, for fixed $b(\bullet)$, the linear system (9.20) for $b(\tau), \tau \in T$ has at most one solution satisfying $b(u \circ v) + b(v \circ u) = 0$ for all $u, v \in T$.

Proof. By assumption on b , the coefficients $b(\tau)$, $\tau \in T \setminus T^*$ are uniquely determined by those for $\tau \in T^*$. The statement is thus obtained in the same way as that for Lemma 9.13 with the only difference that expressions $c(\bullet)b(\sigma)$ and not $c(\sigma)b(\bullet)$ have to be studied. \square

Theorem 9.15 (Chartier, Faou & Murua 2005). *The only symplectic method (as B-series) that conserves the Hamiltonian for arbitrary $H(y)$ is the exact flow of the differential equation.*

Proof. If the method conserves exactly the Hamiltonian, we have (9.20) with $c(\bullet) = 1$ and $c(\tau) = 0$ for all other trees in T^* . By the uniqueness statement of Lemma 9.14 and the symplecticity of the method (Theorem 9.10), we obtain $b(\tau) = 0$ for $|\tau| > 1$. Consequently, no perturbation is permitted in the modified differential equation of the method. \square

A closely related result is given in Ge & Marsden (1988). There, general symplectic methods are considered (not necessarily B-series methods) but a weaker result is obtained (in fact, they assume that the system does not have other conserved quantities than $H(y)$, and it is shown that the numerical flow coincides with the exact flow up to a reparametrization of time).

IX.9.5 Energy Conservation: Examples and Counter-Examples

It is generally believed that symmetric methods applied to reversible Hamiltonian systems (reversible in the sense that $H(-p, q) = H(p, q)$) have the same long-time behaviour as symplectic methods. This is true in many situations of practical interest, and we shall prove this rigorously in Sect. XI.3 for integrable reversible systems. There are, however, interesting counter-examples to this general belief. They are taken from Faou, Hairer & Pham (2004).

Example 9.16. Our first example is a modification of the pendulum equation

$$H(p, q) = \frac{1}{2}p^2 - \cos q + \frac{1}{5} \sin(2q), \quad (9.24)$$

where the additional term $\sin(2q)$ destroys the symmetry in q . The Hamiltonian still satisfies $H(-p, q) = H(p, q)$. We consider initial values $p(0) = 2.5$, $q(0) = 0$ with sufficiently large initial velocity, such that $p(t)$ stays positive for all times and the symmetry $p \leftrightarrow -p$ does not affect the numerical solution. The angle $q(t)$ increases without limit, but the potential is 2π -periodic so that the solution stays on a closed curve of the cylinder $\mathbb{R} \times S^1$.

We apply the 3-stage Lobatto IIIA and IIIB methods to this problem. Figure 9.3 shows the error in the Hamiltonian along the numerical solutions. There is a visible energy drift of size $\mathcal{O}(th^4)$ for the Lobatto IIIB method and no drift can be seen on this scale for the Lobatto IIIA method. To get more insight into its long-time behaviour, we apply the method with the same step size to a much longer time interval, and we plot the error in $H(p_n, q_n) + h^4 H_5(p_n, q_n)$, where the first perturbation term is computed from (9.18) and the linear system (9.20) as

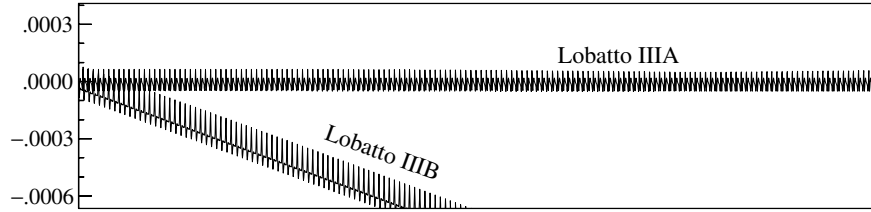


Fig. 9.3. Numerical Hamiltonian of Lobatto methods of order 4 for the perturbed pendulum (9.24); step size $h = 0.2$, integration interval $[0, 500]$

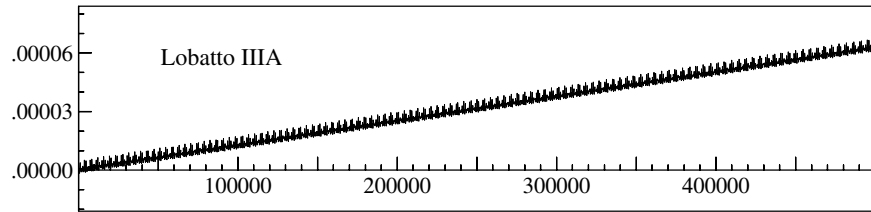


Fig. 9.4. Error in $H(p, q) + h^4 H_5(p, q)$ along the numerical solution of the 3-stage Lobatto IIIA method for the perturbed pendulum (9.24); step size $h = 0.2$, integration interval $[0, 500\,000]$

$$H_5(p, q) = \frac{1}{960} \left(3U^{(4)}(q)p^4 - 2U^{(3)}U'(q)p^2 - (U''(q)p)^2 + U''(q)(U'(q))^2 \right)$$

with the potential $U(q) = -\cos q + 0.2 \sin(2q)$ (see Fig. 9.4). Repeating the same experiment with halved step size shows that there are oscillations with amplitude $\mathcal{O}(h^6)$ and a drift with slope $\mathcal{O}(h^8)$. Consequently, the error in the Hamiltonian for the Lobatto IIIA method behaves on this problem like $\mathcal{O}(h^4 + th^8)$.

Without the term $\sin(2q)$ in (9.24) all symmetric one-step methods nearly conserve the Hamiltonian.

Example 9.17. For polynomial Hamiltonians $H(y)$ of degree at most four, the elementary Hamiltonian corresponding to the tree $\begin{array}{c} \bullet \\ \swarrow \quad \downarrow \quad \searrow \end{array}$ vanishes identically. Therefore, the condition (9.20) need not be considered for this tree, and the remaining three conditions can always be satisfied by the three $c(\tau)$ coefficients. This implies that, for example for the Hénon–Heiles problem

$$H(p_1, p_2, q_1, q_2) = \frac{1}{2}(p_1^2 + p_2^2) + \frac{1}{2}(q_1^2 + q_2^2) + q_1 q_2^2 - \frac{1}{3} q_1^3, \quad (9.25)$$

the leading error term in the numerical Hamiltonian remains bounded by all methods of order four. Numerical experiments indicate that in this case also higher order error terms are bounded by symmetric methods such as Lobatto IIIA and IIIB, even if the initial values are chosen so that the solution is chaotic.

Example 9.18. A concrete mechanical system with two degrees of freedom is described by the Hamiltonian

$$H(p, q) = \frac{1}{2} p^T p + \frac{\omega^2}{2} (\|q\| - 1)^2 + q_2 - \frac{1}{\|q - a\|}. \quad (9.26)$$

It is a model of a planar spring pendulum with exterior forces. The spring has a harmonic potential with frequency ω (Hooke's law). The exterior forces are gravitation and attraction to a mass point situated at a , which has to be chosen so that no symmetry in the q -variables is present.

The numerical experiments, reported by Faou, Hairer & Pham (2004), use $\omega = 2$, $a = (-3, -5)^T$, and initial values for the position $q(0) = (0, 1)^T$ (up-right position), and for the velocity $p(0) = (-1, -0.5)^T$. The pendulum thus turns around the fixed end of the spring which is at the origin.

As for the problem of Example 9.16 one clearly observes a drift for the 3-stage Lobatto IIIB method, and the error in the Hamiltonian behaves like $\mathcal{O}(th^4)$. As predicted by the theory of the preceding section, the dominant error term for the 3-stage Lobatto IIIA method is bounded. There is, however, a drift already in the next term so that the error in the Hamiltonian behaves for this method as $\mathcal{O}(h^4 + th^6)$.

Removing one of the exterior forces (gravitation or attraction to a), the error in the Hamiltonian remains bounded of size $\mathcal{O}(h^4)$ without any drift (even not in higher order terms) for both Lobatto methods.

IX.10 Extension to Partitioned Systems

All results of Sect. IX.9 can be extended to partitioned methods whose discrete flow can be written as a P-series. This includes important geometric integrators such as the symplectic Euler method and the Störmer–Verlet scheme. Interestingly, many of the results have been originally presented and proved for this more general case (see Hairer (1994)).

IX.10.1 P-Series of the Modified Equation

We consider the partitioned system

$$\dot{p} = f(p, q), \quad \dot{q} = g(p, q), \quad (10.1)$$

where, in view of an application to Hamiltonian systems, we use (p, q) instead of (y, z) for the variables. By Theorem III.2.4 all consistent partitioned Runge–Kutta methods can be written as P-series (cf. Definition III.2.1)

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \begin{pmatrix} p_0 \\ q_0 \end{pmatrix} + h \begin{pmatrix} f \\ g \end{pmatrix}_0 + h^2 \begin{pmatrix} a(\mathcal{J})(f_p f) + a(\mathcal{J})(f_q g) \\ a(\mathcal{J})(g_p f) + a(\mathcal{J})(g_q g) \end{pmatrix}_0 + \dots, \quad (10.2)$$

where the subscript 0 indicates an evaluation at the initial value (p_0, q_0) . The first perturbation term of the modified equation (1.1) can therefore be written as

$$\begin{pmatrix} f_2(p, q) \\ g_2(p, q) \end{pmatrix} = \begin{pmatrix} (a(\mathcal{I}) - \frac{1}{2})(f_p f)(p, q) + (a(\mathcal{J}) - \frac{1}{2})(f_q g)(p, q) \\ (a(\mathcal{J}) - \frac{1}{2})(g_p f)(p, q) + (a(\mathcal{I}) - \frac{1}{2})(g_q g)(p, q) \end{pmatrix}$$

and, in general, one finds

$$\begin{pmatrix} f_j(p, q) \\ g_j(p, q) \end{pmatrix} = \begin{pmatrix} \sum_{\tau \in TP_p, |\tau|=j} \frac{b(\tau)}{\sigma(\tau)} F(\tau)(p, q) \\ \sum_{\tau \in TP_q, |\tau|=j} \frac{b(\tau)}{\sigma(\tau)} F(\tau)(p, q) \end{pmatrix}. \quad (10.3)$$

Hence, the modified equation (1.1) is of the form

$$\begin{pmatrix} \dot{\tilde{p}} \\ \dot{\tilde{q}} \end{pmatrix} = \begin{pmatrix} \sum_{\tau \in TP_p} \frac{h^{|\tau|-1}}{\sigma(\tau)} b(\tau) F(\tau)(\tilde{p}, \tilde{q}) \\ \sum_{\tau \in TP_q} \frac{h^{|\tau|-1}}{\sigma(\tau)} b(\tau) F(\tau)(\tilde{p}, \tilde{q}) \end{pmatrix}, \quad (10.4)$$

where $b(\tau) = 1$ for $|\tau| = 1$, $b(\tau) = a(\tau) - \frac{1}{2}$ for $|\tau| = 2$. For $|\tau| > 2$, the coefficients $b(\tau)$ can be obtained recursively from Theorem 10.2 below. The proofs of the following two results are straightforward extensions of those for Lemma 9.1 and Theorem 9.2, and are therefore omitted.

Lemma 10.1 (Lie-Derivative of P-series). *Let $b(\tau)$ (with $b(\emptyset_p) = b(\emptyset_q) = 0$) and $c(\tau)$ be the coefficients of two P-series, and let $(p(t), q(t))$ be a formal solution of the differential equation $h(\dot{p}(t), \dot{q}(t))^T = P(b, (p(t), q(t)))$, i.e., (10.4). The Lie derivative of the function $P(c, (p, q))$ with respect to the vector field $P(b, (p, q))$ is again a P-series*

$$h \frac{d}{dt} P(c, (p(t), q(t))) = P(\partial_b c, (p(t), q(t))).$$

Its coefficients are given by $\partial_b c(\emptyset_p) = \partial_b c(\emptyset_q) = 0$, and for $|\tau| \geq 1$ by

$$\partial_b c(\tau) = \sum_{\theta \in SP(\tau)} c(\theta) b(\tau \setminus \theta), \quad (10.5)$$

where, analogously to (9.5), $SP(\tau)$ denotes the set of splittings of $\tau \in TP$. \square

In formula (10.5), $\emptyset_p \in SP(\tau)$ defines a splitting only if $\tau \in TP_p$, and $\emptyset_q \in SP(\tau)$ only if $\tau \in TP_q$. We therefore have $\partial_b c(\bullet) = c(\emptyset_p)b(\bullet)$, $\partial_b c(\circ) = c(\emptyset_q)b(\circ)$, and as examples for trees of order 3

$$\begin{aligned} \partial_b c(\mathcal{V}) &= c(\emptyset_p)b(\mathcal{V}) + 2c(\mathcal{J})b(\circ), \\ \partial_b c(\mathcal{V}^\circ) &= c(\emptyset_p)b(\mathcal{V}^\circ) + c(\mathcal{I})b(\circ) + c(\mathcal{J})b(\bullet). \end{aligned}$$

Theorem 10.2. *If the method $(p_1, q_1) = \Phi_h(p_0, q_0)$ can be written as (10.2), the modified differential equation is given by (10.4), where the real coefficients $b(\tau)$ are recursively defined by $b(\emptyset_p) = b(\emptyset_q) = 0$, $b(\tau) = 1$ for $|\tau| = 1$, and*

$$b(\tau) = a(\tau) - \sum_{j=2}^{|\tau|} \frac{1}{j!} \partial_b^{j-1} b(\tau) \quad \text{for } \tau \in TP. \quad (10.6)$$

Here, ∂_b^{j-1} denotes the iterate of the Lie derivative ∂_b defined in Lemma 10.1. \square

Example 10.3. The symplectic Euler method











$$p_{n+1} = p_n + hf(p_{n+1}, q_n), \quad q_{n+1} = q_n + hg(p_{n+1}, q_n) \quad (10.7)$$

is a partitioned Runge–Kutta method ($a_{11} = 1$, $\hat{a}_{11} = 0$, $b_1 = \hat{b}_1 = 1$) and can therefore be expressed as a P-series (10.2). From Theorem III.2.4 we get its coefficients:

$$a(\tau) = \begin{cases} 1 & \text{if all vertices (different from the root) are black,} \\ 0 & \text{otherwise.} \end{cases}$$

From Theorem 10.2 we can compute the coefficients $b(\tau)$ of the modified equation (10.4). They are given in Table 10.1 for the trees with a black root. Since $a(\tau)$ does not depend on the colour of the root of τ , the same holds for the coefficients $b(\tau)$. Hence, we do not include the values of $b(\tau)$ for trees with a white root.

Table 10.1. Coefficients $b(\tau)$ of the modified equation for symplectic Euler (10.7)

τ										
$b(\tau)$	1	1/2	-1/2	1/6	-1/3	1/6	1/3	-1/6	-1/6	1/3

We know from Theorem 3.1 that the modified differential equation (10.4) of a symplectic method applied to a Hamiltonian system

$$\dot{p} = -H_q(p, q), \quad \dot{q} = H_p(p, q) \quad (10.8)$$

is again Hamiltonian.

Theorem 10.4. Suppose that for all separable Hamiltonians $H(p, q) = T(p) + U(q)$ the modified vector field (10.4), truncated after an arbitrary power of h , is (locally) Hamiltonian. Then, we have

$$b(u \circ v) + b(v \circ u) = 0 \quad u \in TP_p, v \in TP_q \quad (10.9)$$

for trees, where neighbouring vertices have different colours.

If it is (locally) Hamiltonian for all $H(p, q)$, then (10.9) holds for all $u \in TP_p$, $v \in TP_q$, and additionally we have

$$b(\tau) \text{ is independent of the colour of the root of } \tau \in TP. \quad (10.10)$$

If it is (locally) Hamiltonian for all $H(p, q) = \frac{1}{2}p^T Cp + c^T p + U(q)$ (with symmetric matrix C), then we have

$$b(\circ \circ u) + b(u \circ \circ) = 0, \quad b(u \circ \circ v) - b(v \circ \circ u) = 0 \quad u, v \in TN_p \quad (10.11)$$

(see Sect. VI.7.1 for the definition of TN_p and $u \circ \circ v$).

The proof is the same as for Theorem 9.3 and therefore omitted. \square

IX.10.2 Elementary Hamiltonians

We have already seen in Example 3.4 that the modified Hamiltonian of the symplectic Euler method is composed of expressions such as $H_p H_q$, $H_{pp}(H_q, H_q)$, $H_{pq}(H_q, H_p)$, etc. These will play the role of elementary Hamiltonians for partitioned methods. In the following definition, the elementary differentials $F(\tau)(p, q)$ correspond to the partitioned system $f(p, q) = -H_q(p, q)$, $g(p, q) = H_p(p, q)$.

Definition 10.5. For a given function $H : D \rightarrow \mathbb{R}$ (with open $D \subset \mathbb{R}^d \times \mathbb{R}^d$) and for $\tau \in TP$ we define the *elementary Hamiltonian* $H(\tau) : D \rightarrow \mathbb{R}$ by

$$\begin{aligned} H(\bullet)(p, q) &= H(\circ)(p, q) = H(p, q) \\ H(\tau)(p, q) &= \frac{\partial^{m+l} H(p, q)}{\partial^m p \partial^l q} \left(F(u_1)(p, q), \dots, F(v_1)(p, q), \dots \right) \end{aligned}$$

where $\tau = [u_1, \dots, u_m, v_1, \dots, v_l]_p$ or $\tau = [u_1, \dots, u_m, v_1, \dots, v_l]_q$ with trees $u_i \in TP_p$ and $v_i \in TP_q$.

Examples of elementary Hamiltonians are

$$\begin{aligned} H(\bullet) &= H, & H(\textcircled{\bullet}) &= H_q H_p, \\ H(\textcircled{\text{V}}) &= H_{pp}(H_q, H_q), & H(\textcircled{\text{V}}^\circ) &= -H_{pq}(H_q, H_p), & H(\textcircled{\text{V}}^\circ) &= H_{qq}(H_p, H_p). \end{aligned}$$

We notice that, in contrast to Sect. IX.9.2, non-vanishing elementary Hamiltonians exist for trees with two vertices.

Lemma 10.6. *Elementary Hamiltonians satisfy*

$$H(u \circ v)(p, q) + H(v \circ u)(p, q) = 0 \quad \text{for } u \in TP_p \text{ and } v \in TP_q, \quad (10.12)$$

and they do not depend on the colour of the root.

Proof. The independence of the colour of the root is by definition, and formula (10.12) is proved in the same way as the statement of Lemma 9.6. \square

The conditions (10.9) and (10.10) define relations between the coefficients $b(\tau)$ of a Hamiltonian vector field (10.4). The previous lemma shows analogous relations between elementary Hamiltonians. This motivates the consideration of the following equivalence relation on TP (Hairer 1994).

Definition 10.7. We denote by \sim the smallest *equivalence relation* on TP which satisfies the two properties

- $u \sim v$ if u and v are identical with the exception of the colour of the root;
- $u \circ v \sim v \circ u$ for $u \in TP_p$ and $v \in TP_q$.



Fig. 10.1. Groups of equivalent trees of orders up to three

Equivalent trees of orders up to three are grouped together in Fig. 10.1. We can change the colour of the root, and we can move the root to a neighbouring vertex if it has the opposite colour.

In the case of separable Hamiltonians, one has to consider only trees for which neighbouring vertices have different colours. This implies that the first condition of Definition 10.7 is empty. The second condition means that the root can be moved arbitrarily in the tree without changing the equivalence class. For this special situation, equivalence classes have been considered already by Abia & Sanz-Serna (1993) and are named “bicolour (unrooted) trees”.

Similar to (9.14) we select representatives from the equivalence class as follows: we fix a total ordering on the set TP that (i) respects the number of vertices, and (ii) is such that no tree is between trees that differ only in the colour of the root. The ordering of Fig. 10.1 is such a possible choice. We then define

$$TP^* = \left\{ \bullet, \circ \right\} \cup \left\{ \tau \in TP \mid \begin{array}{l} \tau \text{ cannot be written as } \tau = u \circ v \text{ with } u < v, \\ \text{also not if the colour of the root is changed.} \end{array} \right\}. \quad (10.13)$$

We further let $TP_p^* = TP^* \cap TP_p$ and $TP_q^* = TP^* \cap TP_q$.

Lemma 10.8. *For a tree $\tau \in TP^*$ we have*

$$\begin{aligned} -\frac{\partial H(\tau)}{\partial q}(p, q) &= \sigma(\tau) \sum_{\theta \sim \tau, \theta \in TP_p} \frac{(-1)^{\kappa(\tau, \theta)}}{\sigma(\theta)} F(\theta)(p, q), \\ \frac{\partial H(\tau)}{\partial p}(p, q) &= \sigma(\tau) \sum_{\theta \sim \tau, \theta \in TP_q} \frac{(-1)^{\kappa(\tau, \theta)}}{\sigma(\theta)} F(\theta)(p, q), \end{aligned} \quad (10.14)$$

where $\kappa(\tau, \theta)$ is the number of root changes that are necessary to obtain θ from τ .

The proof is the same as for Lemma 9.7 and therefore omitted. \square

We are now able to give the main result of this section.

Theorem 10.9. *Consider a numerical method that can be written as a P -series (10.2), and that is symplectic for every Hamiltonian (10.8). Its modified differential equation is then Hamiltonian with*

$$\tilde{H}(p, q) = H_1(p, q) + h H_2(p, q) + h^2 H_3(p, q) + \dots,$$

where

$$H_j(p, q) = \sum_{\tau \in TP_p^*, |\tau|=j} \frac{b(\tau)}{\sigma(\tau)} H(\tau)(p, q), \quad (10.15)$$

and the coefficients $b(\tau)$ are those of Theorem 10.2. Notice that $H_j(p, q)$ from (10.15) is independent of whether we sum over trees in TP_p^* or TP_q^* .

Proof. This is the same as for Theorem 9.8. \square

If the method (10.2) is known to be symplectic for separable Hamiltonians only, and if it is applied to $H(p, q) = T(p) + U(q)$, the statement of Theorem 10.9 is still valid. In this situation $H(\tau)(p, q)$ vanishes if a vertex of τ has sons with different colour (it then contains a factor $H_{pq\dots} = 0$).

Example 10.10. Consider the 2-stage Lobatto IIIA - IIIB pair (cf. Table II.2.1), which is the natural extension of the Störmer–Verlet scheme to non-separable problems. We compute the coefficients $a(\tau)$ from Theorem III.2.4, and $b(\tau)$ from Theorem 10.2. The result is given in Table 10.2. Notice that $a(\tau)$ and $b(\tau)$ are both independent of the colour of the root. Theorem 10.9 then yields

$$\tilde{H} = H + \frac{h^2}{24} (2H_{pp}H_q^2 - H_{qq}H_p^2 + 2H_{pq}H_qH_p) + \dots \quad (10.16)$$

for the modified Hamiltonian. Since the method is symmetric, \tilde{H} is in even powers of h . The next non-vanishing term requires the consideration of trees up to order 5.

Table 10.2. Coefficients $a(\tau)$ and $b(\tau)$ for the Störmer–Verlet scheme (Table II.2.1)

τ	\bullet									
$a(\tau)$	1	1/2	1/2	1/2	1/4	1/4	1/4	1/4	0	1/4
$b(\tau)$	1	0	0	1/6	-1/12	-1/12	1/12	1/12	-1/6	1/12

Remark 9.9, the characterization of symplectic vector fields (10.4), and the results of Sect. IX.9.4 can be extended to the case of (partitioned) P-series. We re-nounce of giving all the details here.

IX.11 Exercises

1. Change the Maple program of Example 1.1 in such a way that the modified equations for the implicit Euler method, the implicit midpoint rule, or the trapezoidal rule are obtained. Observe that for symmetric methods one gets expansions in even powers of h .
2. Write a short Maple program which, for simple methods such as the symplectic Euler method, computes some terms of the modified equation for a two-dimensional system $\dot{p} = f(p, q)$, $\dot{q} = g(p, q)$. Check the modified equations of Example 1.3.
3. Prove that the modified equation of the Störmer–Verlet scheme (I.1.15) applied to $\ddot{y} = g(y)$ is a second order differential equation of the form $\ddot{\tilde{y}} = g_h(\tilde{y}, \dot{\tilde{y}})$ with initial values given by $\tilde{y}(0) = y_0$ and $\dot{\tilde{y}}(0)$ such that $\tilde{y}(h) = y_1$ holds.

Hint. Taylor expansion shows that for a smooth function $\tilde{y}(t)$ satisfying $\tilde{y}(t) = y_n$ we have

$$\left(1 + \frac{h^2}{12} D^2 + \frac{h^4}{360} D^4 + \dots\right) \ddot{\tilde{y}}(t) = g(\tilde{y}(t)),$$

where D represents differentiation with respect to time.

Warning. In general, we do not have that $\tilde{y}(t_n) = \dot{y}_n$.

4. Prove that for ρ -reversible differential equations the elementary differentials satisfy

$$F(\tau)(\rho y) = (-1)^{|\tau|} \rho F(\tau)(y).$$

Use this to give an alternative proof of Theorem 2.3 for the case that the method is symmetric and can be expressed as a B -series.

5. Find a first integral of the truncated modified equation for the symplectic Euler method and the Lotka–Volterra problem (Example 1.3).

Hint. With the transformation $p = \exp P$, $q = \exp Q$ you will get a Hamiltonian system.

Result. $\tilde{I}(p, q) = I(p, q) - h((p + q)^2 - 8p - 10q + 2 \ln p + 8 \ln q)/4$.

6. (Field & Nijhoff 2003). Apply the symplectic Euler method to the system with Hamiltonian $H(p, q) = \ln(\alpha + p) + \ln(\beta + q)$. Compute the modified Hamiltonian and prove that the series converges for sufficiently small step sizes.

Hint. The method conserves exactly $I(p, q) = (\alpha + p)(\beta + q)$. Find linear two-term recursions for $\{p_n\}$ and $\{q_n\}$, and use the ideas of Example 1.4. *Result.*

$$\tilde{H}(p, q) = H(p, q) - \sum_{k \geq 1} \frac{h^k I(p, q)^{-k}}{k(k+1)}.$$

7. Compute $\partial_b c(\tau)$ for the tree $\tau = [[\tau], \tau]$ of order 4.
 8. For the implicit midpoint rule compute the coefficients $a(\tau)$ of the expansion (9.1), and also a few coefficients $b(\tau)$ of the modified equation.
Result. $a(\tau) = 2^{1-|\tau|}$, $b(\bullet) = 1$, $b(\text{hook}) = 0$, $b(\tau) = a(\tau) - 1/\gamma(\tau)$ for $|\tau| = 3$.

9. Check the formulas of Table 9.1.
 10. Consider a differential equation $\dot{y} = f(y)$ with a divergence-free vector field, and apply a volume-preserving integrator. Show that every truncation of the modified equation has again a divergence-free vector field.
Hint. Adapt the proof by induction of Theorems 2.3 and 3.1.
 11. Consider explicit 2-stage Runge–Kutta methods of order 2, applied to the pendulum problem $\dot{q} = p$, $\dot{p} = -\sin q$. With the help of Exercise 2 compute $f_3(p, q)$ of the modified differential equation. Is there a choice of the free parameter c_2 , such that $f_3(p, q)$ is a Hamiltonian vector field?
 12. Find at least two linear transformations ρ for which the Kepler problem (I.2.2), written as a first order system, is ρ -reversible.
 13. Consider the Kepler problem (I.2.2), written as a Hamiltonian system (I.1.10). Find constants M and R such that (7.2) holds for all $(p, q) \in \mathbb{R}^4$ satisfying

$$\|p\| \leq 2 \quad \text{and} \quad 0.8 \leq \|q\| \leq 1.2.$$

14. (McLachlan & Zanna 2005). Consider the RATTLE method (Algorithm VII.5.1) applied to the Euler equations (VII.5.10) of the free rigid body, written as $\dot{y} = f(y)$. Prove that the modified differential equation is of the form

$$\dot{y} = (1 + h^2 s_2(y) + h^4 s_4(y) + \dots) f(y), \quad (11.1)$$

where the scalar functions $s_k(y)$ depend on y only via the Casimir function $C(y) = y_1^2 + y_2^2 + y_3^2$ and the Hamiltonian $H(y) = \frac{1}{2}(y_1^2/I_1 + y_2^2/I_2 + y_3^2/I_3)$. Consequently, all $s_k(y)$ are constant along solutions of the Euler equations.

Hint. Since $C(y)$ and $H(y)$ are exactly conserved by the numerical method (see Sect. VII.5.3), the modified equation is a time transformation of the original system. The special form of the functions $s_k(y)$ follows from the fact that RATTLE is a Poisson integrator (Theorem VII.5.11) and from a transformation to canonical form as in Theorem 3.5.

15. (Murua 1999). Let $\Phi_h(y) = B(a, y)$ be given by a B-series and denote with $b(\tau)$ the coefficients of the corresponding modified differential equation, cf. formula (9.4). Prove that the coefficients of the n th iterate $\Phi_h^n(y) = B(a^n, y)$ satisfy

$$a^n(\tau) = n b(\tau) + n^2 c(\tau, n) \quad \text{for } \tau \in T,$$

where $c(\tau, n)$ is a polynomial of degree $|\tau| - 2$ in n .

Hint. This follows from the Taylor series $\tilde{y}(nh) = \tilde{y}(0) + nh\tilde{y}'(0) + \dots$ for the solution of the modified differential equation.

16. With the help of Exercise 15, give an alternative proof of Theorem 9.3.
Hint. If $B(a, y)$ is symplectic, also $B(a^n, y)$ is symplectic and its coefficients thus satisfy (VI.7.4).
17. (Murua 1997). Find a one-to-one correspondence between the equivalence classes of TP (corresponding to \sim of Definition 10.7) and *oriented free trees* (i.e., trees without a distinguished vertex (root), but with oriented edges), see Fig. 11.1.

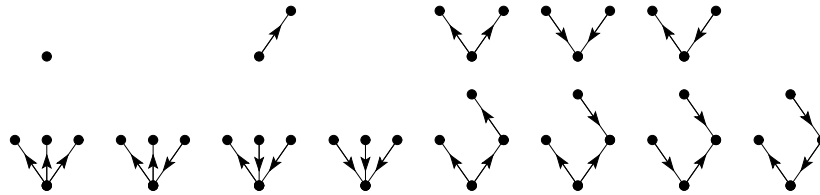


Fig. 11.1. Oriented free trees up to order four

Chapter X.

Hamiltonian Perturbation Theory and Symplectic Integrators

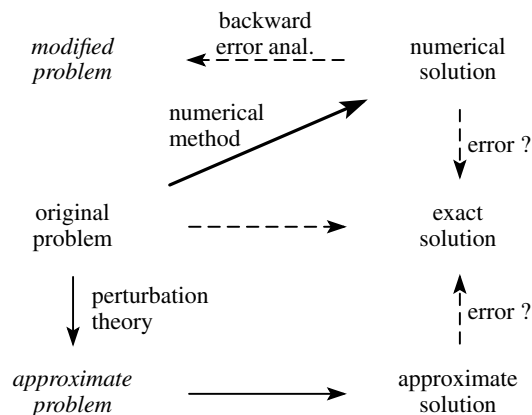
Perturbation theory is in fact an outgrowth of the necessity to determine the orbits with ever greater accuracy. This problem can be solved today, but in what is for the theoretician a rather disappointing way. With modern calculating machines, one is now able to compute directly results even more accurately than those provided by perturbation theory.

(J. Moser 1978)

... allows computer prediction of planetary positions far more accurate (by brute computation) than anything provided by classical perturbation theory. In a very real sense, one of the most exalted of human endeavors, going back to the priests of Babylon and before, has been taken over by the machine.

(S. Sternberg 1969)

In this chapter we study the long-time behaviour of symplectic integrators, combining backward error analysis and the perturbation theory of integrable Hamiltonian systems.



During the 18th and 19th centuries, scientists struggled for the integration of complicated problems of dynamics, with the main aim of solving them analytically by “quadrature”. But only few problems could be treated successfully in this way. In cases where the original problem could not be solved, much effort was put into re-

placing it by an integrable *approximate problem*, by using and developing perturbation theory. Thereby, a rich arsenal of very ingenious theories has been discovered since the 19th century.

In the 1960s and 1970s, the enormous progress of “calculating machines” and numerical software allowed many of the original problems to be solved with extreme accuracy, so that for the first time numerical integration methods superseded analytical perturbation methods in the computations of celestial mechanics (see the above citations). Since then, the further increase in computing speed has allowed problems to be treated on larger and larger time scales, where huge amounts of errors are accumulated and need to be understood and controlled. In the spirit of backward error analysis, these numerical errors are interpreted as those of a *modified problem*, for the study of which perturbation theory is once again the appropriate tool.

X.1 Completely Integrable Hamiltonian Systems

Integrable Hamiltonian systems were originally of interest because their equations of motion can be solved analytically. Their interest in the present context lies in the fact that their flow is simply uniform motion on a Cartesian product of circles and straight lines in suitable coordinates, and that many physical systems can be viewed as perturbations of integrable systems.

X.1.1 Local Integration by Quadrature

M. Liouville a fait voir qu'il fallait que toutes les combinaisons (α, β) des intégrales trouvées fussent nulles. (E. Bour 1855)

One of the great dreams of 18th and 19th century analytical mechanics was to solve the equations of motion of mechanical systems by “quadrature”, that is, using only evaluations and inversions of functions and calculating integrals of known functions. In this spirit, Newton’s (1687) equations of motion of Kepler’s two-body problem were solved by Joh. Bernoulli (1710) and Newton (1713), see Sect. I.2.2. Euler’s (1760) solution of the problem of the attraction of a particle by two fixed centres, and Lagrange’s (1766) study of motion of a particle in a field with one attracting centre and under an additional constant force were among the important achievements of the 18th century. The three-body problem, however, resisted all efforts aiming at an integration by quadrature, and though it continued to do so, this problem spurred the development of extremely useful mathematical theories of a much wider scope throughout the 19th century, from Poisson to Poincaré via Hamilton, Jacobi, Liouville, to name but a few of the most eminent mathematicians contributing to analytical mechanics.

Consider the Hamiltonian system

$$\dot{p} = -\frac{\partial H}{\partial q}(p, q), \quad \dot{q} = \frac{\partial H}{\partial p}(p, q), \quad (1.1)$$

with d degrees of freedom: $(p, q) \in \mathbb{R}^d \times \mathbb{R}^d$. We try to find a symplectic transformation $(p, q) \mapsto (x, y)$, such that the system has a more amenable form in the new coordinates. In particular, this is the case if the Hamiltonian expressed in the new variables,

$$H(p, q) = K(x) , \quad (1.2)$$

does not depend on y . Since $\frac{\partial K}{\partial y} \equiv 0$, the transformed system then becomes (recall the conservation of the Hamiltonian form of the differential equations under symplectic transformations, Theorem VI.2.8)

$$\dot{x} = 0 , \quad \dot{y} = \omega(x) , \quad (1.3)$$

with $\omega(x) = \frac{\partial K}{\partial x}(x)$. This is readily integrated:

$$x(t) = x_0 , \quad y(t) = y_0 + \omega(x_0)t .$$

As we recall from Sect. VI.5, a symplectic transformation $(p, q) \mapsto (x, y)$ can be constructed via a *generating function* $S(x, q)$ by the equations

$$y = \frac{\partial S}{\partial x}(x, q) , \quad p = \frac{\partial S}{\partial q}(x, q) . \quad (1.4)$$

If (p_0, q_0) and (x_0, y_0) are related by (1.4), and if $\partial^2 S / \partial x \partial q$ is invertible at (x_0, q_0) , then the equations (1.4) define a symplectic transformation between neighbourhoods of (p_0, q_0) and (x_0, y_0) .

The equation (1.2) together with the second equation of (1.4) give a partial differential equation for S , the *Hamilton–Jacobi equation*

$$H\left(\frac{\partial S}{\partial q}(x, q), q\right) = K(x) .$$

If $S(x, q)$ is a solution of such an equation (for some function K), then (1.3) shows that $x_i = F_i(p, q)$ ($i = 1, \dots, d$) as given implicitly by the second equation of (1.4), are first integrals of the Hamiltonian system (1.1). Moreover, these functions F_i are *in involution*, which means that their Poisson brackets vanish pairwise:

$$\{F_i, F_j\} = 0, \quad i, j = 1, \dots, d .$$

This is an immediate consequence of the definition $\{F, G\} = \nabla F^T J^{-1} \nabla G$ of the Poisson bracket and of the symplecticity of the transformation (the left upper block of J^{-1} is 0).

Conversely, it was realized by Bour (1855) and Liouville (1855) that a Hamiltonian system having d first integrals in involution can *locally* be transformed to the form (1.3) by “quadrature”. This observation is based on the following completion result and its proof.

Lemma 1.1 (Liouville Lemma). *Let F_1, \dots, F_d be smooth real-valued functions, defined in a neighbourhood of $(p_0, q_0) \in \mathbb{R}^d \times \mathbb{R}^d$. Suppose that these functions are in involution (i.e., all Poisson brackets $\{F_i, F_j\} = 0$), and that their gradients are linearly independent at (p_0, q_0) . Then, there exist smooth functions G_1, \dots, G_d , defined on some neighbourhood of (p_0, q_0) , such that*

$$(F_1, \dots, F_d, G_1, \dots, G_d) : (p, q) \mapsto (x, y) \text{ is a symplectic transformation.}$$

Proof. Let $F = (F_1, \dots, F_d)^T$. The linear independence of the gradients ∇F_i implies that there are d columns of the $d \times 2d$ Jacobian $\partial F / \partial(p, q)$ that form an invertible $d \times d$ submatrix. After some suitable symplectic transformations (see Exercise 1) we may assume without loss of generality that $F_p = \partial F / \partial p$ is invertible. By the implicit function theorem, we can then locally solve $x = F(p, q)$ for p :

$$p = P(x, q) \quad \text{with partial derivatives} \quad P_x = F_p^{-1}, \quad P_q = -F_p^{-1} F_q.$$

The condition that the F_i are in involution, reads in matrix notation

$$F_p F_q^T - F_q F_p^T = 0.$$

Multiplying this equation with F_p^{-1} from the left and with F_p^{-T} from the right, we obtain

$$-P_q^T + P_q = 0,$$

so that $P_q = \partial P / \partial q$ is symmetric. By the Integrability Lemma VI.2.7, $P(x, q)$ is thus locally the gradient with respect to q of some function $S(x, q)$ (which is constructed by quadrature). Moreover, $\frac{\partial^2 S}{\partial x \partial q} = P_x = F_p^{-1}$ is invertible. The equations (1.4) define a symplectic transformation $(p, q) \mapsto (x, y)$, and by construction $x = F(p, q)$. \square

If, in a Hamiltonian system with d degrees of freedom, we can find d independent first integrals in involution $H = F_1, F_2, \dots, F_d$, then Lemma 1.1 yields a symplectic change of coordinates, constructed by quadrature, which transforms (1.1) locally to (1.2) with $K(x_1, \dots, x_d) = x_1$.

Example 1.2. Consider the Hamiltonian of motion in a central field,

$$H = \frac{1}{2}(p_1^2 + p_2^2) + V(r) \quad \text{for} \quad r = \sqrt{q_1^2 + q_2^2},$$

with a potential $V(r)$ that is defined and smooth for $r > 0$. The Kepler problem corresponds to the special case $V(r) = -1/r$, and the perturbed Kepler problem to $V(r) = -1/r - \mu/(3r^3)$. Changing to polar coordinates (see Example VI.5.2)

$$\begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} r \cos \varphi \\ r \sin \varphi \end{pmatrix}, \quad \begin{pmatrix} p_r \\ p_\varphi \end{pmatrix} = \begin{pmatrix} \cos \varphi & \sin \varphi \\ -r \sin \varphi & r \cos \varphi \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}, \quad (1.5)$$

this becomes

$$H(p_r, p_\varphi, r, \varphi) = \frac{1}{2} \left(p_r^2 + \frac{p_\varphi^2}{r^2} \right) + V(r) .$$

The system has the angular momentum $L = p_\varphi$ as a first integral, since H does not depend on φ . Clearly, $\{H, L\} = 0$ everywhere. The gradients of H and L are linearly independent unless both $p_r = 0$ and $p_\varphi^2 = r^3 V'(r)$. By inserting $p_\varphi^2 = 2r^2(H - V(r))$ and eliminating r this becomes a condition of the form $\alpha(H, L) = 0$, which for the Kepler problem reads explicitly $L^2(1 + 2HL^2) = 0$. The conditions of Lemma 1.1 are thus satisfied on the domain

$$M = \{(p_r, p_\varphi, r, \varphi) ; r > 0, \alpha(H, L) \neq 0\} .$$

The equations $x_1 = H = \frac{1}{2}(p_r^2 + p_\varphi^2/r^2) + V(r)$, $x_2 = L = p_\varphi$ can be solved for

$$p_r = \pm \sqrt{2(H - V(r)) - L^2/r^2} , \quad p_\varphi = L ,$$

and $p_r = \partial S / \partial r$, $p_\varphi = \partial S / \partial \varphi$ with

$$S(H, L, r, \varphi) = L\varphi \pm \int_{r_0}^r \sqrt{2(H - V(\rho)) - L^2/\rho^2} d\rho .$$

The conjugate variables are

$$\begin{aligned} y_1 &= \frac{\partial S}{\partial H} = \pm \int_{r_0}^r \frac{1}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}} d\rho , \\ y_2 &= \frac{\partial S}{\partial L} = \varphi \mp \int_{r_0}^r \frac{L/\rho^2}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}} d\rho . \end{aligned} \quad (1.6)$$

This defines (locally) the transformation $(p_r, p_\varphi, r, \varphi) \mapsto (x_1, x_2, y_1, y_2)$. In these variables, the equations of motion read $\dot{x}_1 = 0$, $\dot{x}_2 = 0$, $\dot{y}_1 = 1$, $\dot{y}_2 = 0$. Over any time interval where $p_r(t)$ does not change sign, solutions therefore satisfy

$$\begin{aligned} t_1 - t_0 &= \pm \int_{r(t_0)}^{r(t_1)} \frac{1}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}} d\rho , \\ \varphi(t_1) - \varphi(t_0) &= \pm \int_{r(t_0)}^{r(t_1)} \frac{L/\rho^2}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}} d\rho . \end{aligned} \quad (1.7)$$

X.1.2 Completely Integrable Systems

Lemma 1.1 appears as a powerful tool for an explicit solution by quadrature. However, because of its purely local nature this lemma does not tell us anything about the dynamics of the system. This was not a concern at Liouville's time, but the first rigorous non-integrability results by Poincaré (1892) put a definite end to the hope of being eventually able to construct explicit analytic solutions of most equations of motion by quadrature, and shifted the interest to understanding the *global*, qualitative behaviour of dynamical systems.

Lemma 1.1 can be globalized by a procedure similar to analytic continuation if the conditions of the following definition are satisfied.

Definition 1.3. A Hamiltonian system with Hamiltonian $H : M \rightarrow \mathbb{R}$ (M an open subset of $\mathbb{R}^d \times \mathbb{R}^d$) is called *completely integrable* if there exist smooth functions $F_1 = H, F_2, \dots, F_d : M \rightarrow \mathbb{R}$ with the following properties:

- 1) F_1, \dots, F_d are in involution (i.e., all $\{F_i, F_j\} = 0$) on M .
- 2) The gradients of F_1, \dots, F_d are linearly independent at every point of M .
- 3) The solution trajectories of the Hamiltonian systems with Hamiltonian F_i ($i = 1, \dots, d$) exist for all times and remain in M .

Obviously, all the Hamiltonian systems with Hamiltonian F_i ($i = 1, \dots, d$) are then completely integrable, and so there will be no mathematical reason to further distinguish $H = F_1$. We note that condition (1) of Definition 1.3 implies that all F_j are first integrals of the Hamiltonian system with Hamiltonian F_i , and that the flows $\varphi_t^{[i]}$ of these Hamiltonian systems commute: $\varphi_t^{[i]} \circ \varphi_s^{[j]} = \varphi_s^{[j]} \circ \varphi_t^{[i]}$ for all i, j and all $t, s \in \mathbb{R}$; see Lemma VII.3.2.

For $x = (x_i) \in \mathbb{R}^d$ we define the level set

$$M_x = \{(p, q) \in M ; F_i(p, q) = x_i \text{ for } i = 1, \dots, d\}. \quad (1.8)$$

Theorem 1.4. Suppose that $F_1, \dots, F_d : M \rightarrow \mathbb{R}$ satisfy the conditions of Definition 1.3. Assume that M_x is connected (and non-empty) for all x in a neighbourhood of $x_0 \in \mathbb{R}^d$. Then, on some neighbourhood B of x_0 , there exists a symplectic and surjective mapping

$$e : B \times \mathbb{R}^d \rightarrow \bigcup_{x \in B} M_x : (x, y) \mapsto (p, q) \in M_x$$

that linearizes, for all $i = 1, \dots, d$, the flow $\varphi_t^{[i]}$ of the system with Hamiltonian F_i :

$$\text{if } (p, q) = e(x, y), \quad \text{then } \varphi_t^{[i]}(p, q) = e(x, y + te_i), \quad (1.9)$$

where $e_i = (0, \dots, 1, \dots, 0)^T$ is the i th unit vector of \mathbb{R}^d .

Since e is symplectic, e is a local diffeomorphism. Its local inverse is a transformation as constructed in Lemma 1.1. However, (p, q) can have countably many discretely lying pre-images (x, y) , so that e^{-1} becomes a multi-valued function. The situation is analogous to that of the complex exponential and logarithm. The following example illustrates that this analogy is not incidental.

Example 1.5. Consider the harmonic oscillator, i.e., $d = 1$ and $H(p, q) = \frac{1}{2}(p^2 + q^2)$. For $x = \frac{1}{2}r^2$, we have $e(x, y) = (r \cos y, r \sin y)$.

Proof of Theorem 1.4. We fix $(p_0, q_0) \in M_{x_0}$, and in a neighbourhood U of (p_0, q_0) we consider a symplectic transformation

$$\ell = (F_1, \dots, F_d, G_1, \dots, G_d) : (p, q) \mapsto (x, y)$$

as constructed in Lemma 1.1. We have $\ell(p_0, q_0) = (x_0, y_0)$ where we may assume $y_0 = 0$. To every $v = (v_i) \in \mathbb{R}^d$ we associate the Hamiltonian

$$F_v = v_1 F_1 + \dots + v_d F_d$$

and note that, because of the commutativity of the flows $\varphi_t^{[i]}$, the flow of the system with Hamiltonian F_v equals

$$\varphi_{tv} = \varphi_{tv_1}^{[1]} \circ \dots \circ \varphi_{tv_d}^{[d]}.$$

In the neighbourhood U of (p_0, q_0) , the system with Hamiltonian F_v is transformed under the symplectic mapping ℓ to

$$\dot{x} = 0, \quad \dot{y} = v.$$

Hence, the following diagram commutes for $(p, q) \in U$ and for sufficiently small tv :

$$\begin{array}{ccc} (p, q) & \longrightarrow & \varphi_{tv}(p, q) \\ \downarrow \ell & & \uparrow \ell^{-1} \\ (x, y) & \longrightarrow & (x, y + tv) \end{array} \quad (1.10)$$

We now construct e by extending this diagram to arbitrary tv :

$$\begin{array}{ccc} (p, q) & \longrightarrow & \varphi_y(p, q) \\ \uparrow \ell^{-1} & & \\ (x, 0) & \longleftarrow & (x, y) \end{array} \quad (1.11)$$

That is, we define on $B \times \mathbb{R}^d$ (with B a neighbourhood of x_0 on which $\ell^{-1}(x, 0)$ is defined)

$$e(x, y) = \varphi_y(\ell^{-1}(x, 0)).$$

For (x, y) near some fixed (\hat{x}, \hat{y}) , we have by (1.10) with $y - \hat{y}$ and \hat{y} instead of y and tv that

$$e(x, y) = \varphi_{\hat{y}}(\ell^{-1}(x, y - \hat{y})),$$

which shows that e is symplectic, being locally the composition of symplectic transformations. The property (1.9) is obvious from the definition of e and from the commutativity of the flows $\varphi_t^{[i]}$. Since $\ell^{-1}(x, 0) \in M_x$ and M_x is invariant under the flows $\varphi_t^{[i]}$, we have $e(x, y) \in M_x$ for all (x, y) .

It remains to show that $e : \{x\} \times \mathbb{R}^d \rightarrow M_x$ is surjective for every x near x_0 . Let (\hat{p}, \hat{q}) be an arbitrary point on M_x . By assumption, there exists a path on M_x connecting $\ell^{-1}(x, 0)$ and (\hat{p}, \hat{q}) . Moreover, by (1.10) and by the compactness of the path, there is a $\delta > 0$ such that, for every (p, q) on this path, the mapping $y \mapsto \varphi_y(p, q)$ is a diffeomorphism between the ball $\|y\| < \delta$ and a neighbourhood of (p, q) on M_x . Therefore, (\hat{p}, \hat{q}) can be reached from $\ell^{-1}(x, 0)$ by a finite composition of maps:

$$(\hat{p}, \hat{q}) = \varphi_{y^{(m)}} \circ \dots \circ \varphi_{y^{(1)}}(\ell^{-1}(x, 0)) = \varphi_{\hat{y}}(\ell^{-1}(x, 0)) = e(x, \hat{y}),$$

where $\hat{y} = y^{(1)} + \dots + y^{(m)}$ once again by the commutativity of the flows $\varphi_t^{[i]}$. \square

Illustration of the Liouville Transform. We illustrate the above construction at a simple example, the pendulum (I.1.12) with Hamiltonian $H = p^2/2 - \cos q$. The first coordinate is $x = H(p, q)$, a first integral. The second coordinate y is, following (1.11), the time t which is necessary to reach the point (p, q) from an initial line, which we assume at $q = 0$. Then we have (Fig. 1.1 left) $dp dq = dH dt$ (because

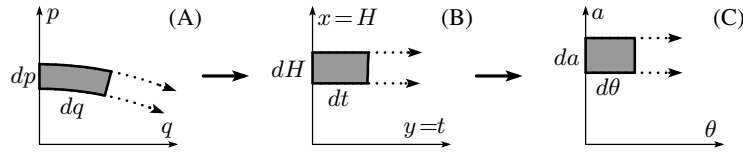


Fig. 1.1. Liouville and action-angle coordinate transforms

of $dq = H_p dt$ and $dH = H_p dp$). We see again that we have area preservation, because the symplecticity of the flow preserves this property for all times. This symplectic change of coordinates $(p, q) \mapsto (x, y)$ is illustrated in Fig. 1.2, which transforms the problem (A) to a much simpler form (B) with uniform horizontal movement.

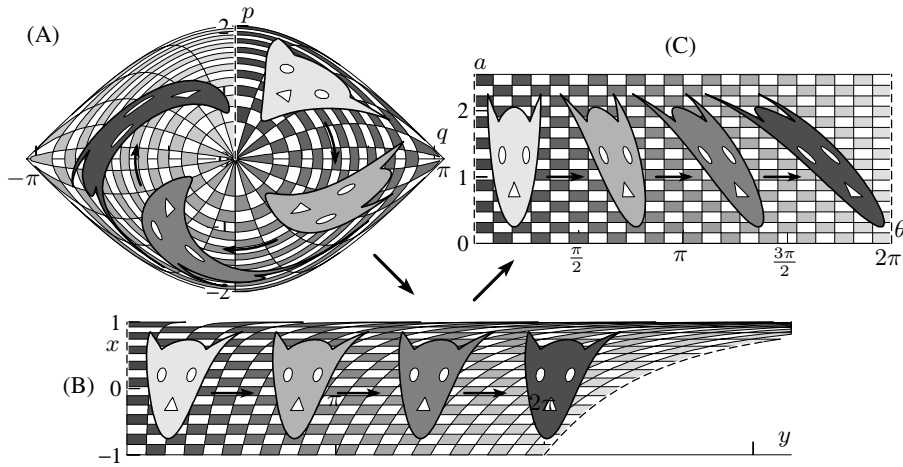


Fig. 1.2. Liouville and action-angle coordinates illustrated at the pendulum problem

We are not yet completely satisfied, however, because the orbits have periods $g = g(H)$ which are not all the same. We therefore append a *second* transform by putting $\theta = \frac{2\pi}{g} \cdot t$ (see picture (C) in Fig. 1.1 and Fig. 1.2), which forces all periods into a Procrustean bed of length 2π . Area preservation $da d\theta = dH dt$ now requires that $2\pi da = g(H) dH$, which is a differential equation between a and H . The new coordinates (a, θ) are the *action-angle variables* and we see that they transform the phase space into $D \times \mathbb{T}^1$ where $D \subset \mathbb{R}^1$. We again have horizontal movement, but this time the speed depends on a . The general existence for completely integrable systems will be proved in Theorem 1.6 below.

X.1.3 Action-Angle Variables

We show here that, under the hypotheses of Liouville's theorem, we can find symplectic coordinates (\mathbf{I}, φ) such that the first integrals \mathbf{F} depend only on \mathbf{I} , and φ are angular coordinates on the torus $M_{\mathbf{f}}$.

(V.I. Arnold 1989, p. 279)

We are now in the position to prove the main result of this section, which establishes a symplectic change of coordinates to the so-called *action-angle variables*, such that d first integrals of a completely integrable system depend only on the actions, and the angles are defined globally mod 2π (provided the level sets of the first integrals are compact). This is known as the Arnold–Liouville theorem; cf. Arnold (1963, 1989), Arnold, Kozlov & Neishtadt (1997; Ch. 4, Sect. 2.1), Jost (1968). Here and in the following,

$$\mathbb{T}^d = \mathbb{R}^d / 2\pi\mathbb{Z}^d = \{(\theta_1 \bmod 2\pi, \dots, \theta_d \bmod 2\pi) ; \theta_i \in \mathbb{R}\}$$

denotes the standard d -dimensional torus.

Theorem 1.6 (Arnold–Liouville Theorem). *Let $F_1, \dots, F_d : M \rightarrow \mathbb{R}$ be first integrals of a completely integrable system as in Definition 1.3. Suppose that the level sets M_x (see (1.8)) are compact and connected for all x in a neighbourhood of $x_0 \in \mathbb{R}^d$. Then, there are neighbourhoods B of x_0 and D of 0 in \mathbb{R}^d such that the following holds:*

(i) *For every $x \in B$, the level set M_x is a d -dimensional torus that is invariant under the flow of the system with Hamiltonian F_i ($i = 1, \dots, d$).*

(ii) *There exists a bijective symplectic transformation*

$$\psi : D \times \mathbb{T}^d \rightarrow \bigcup_{x \in B} M_x \subset \mathbb{R}^d \times \mathbb{R}^d : (a, \theta) \mapsto (p, q)$$

such that $(F_i \circ \psi)(a, \theta)$ depends only on a , i.e.,

$$F_i(p, q) = f_i(a) \quad \text{for } (p, q) = \psi(a, \theta) \quad (i = 1, \dots, d)$$

with functions $f_i : D \rightarrow \mathbb{R}$.

The variables $(a, \theta) = (a_1, \dots, a_d, \theta_1 \bmod 2\pi, \dots, \theta_d \bmod 2\pi)$ are called *action-angle variables*.

Remark 1.7. If the level sets M_x are not compact, then the proof of Theorem 1.6 shows that M_x is diffeomorphic to a Cartesian product of circles and straight lines $\mathbb{T}^k \times \mathbb{R}^{d-k}$ for some $k < d$, and there is a bijective symplectic transformation $(a, \theta) \mapsto (p, q)$ between $D \times (\mathbb{T}^k \times \mathbb{R}^{d-k})$ and a neighbourhood $\bigcup \{M_x : x \in B\}$ of M_{x_0} such that the first integrals again depend only on a .

Remark 1.8. If the Hamiltonian is real-analytic, then the proof shows that also the transformation to action-angle variables is real-analytic.

Proof of Theorem 1.6. (a) We return to Theorem 1.4. For $x \in B$, we consider the set

$$\Gamma_x = \{y \in \mathbb{R}^d; e(x, y) = e(x, 0)\}.$$

Since e is locally a diffeomorphism, for every fixed $y_0 \in \Gamma_{x_0}$ there exists a unique smooth function η defined on a neighbourhood of x_0 , such that $\eta(x_0) = y_0$ and $\eta(x) \in \Gamma_x$ for x near x_0 . In particular, Γ_x is a discrete subset of \mathbb{R}^d . By (1.9), for $y \in \Gamma_x$ we have $e(x, y + v) = e(x, y)$ for all $v \in \mathbb{R}^d$. Therefore, Γ_x is a subgroup of \mathbb{R}^d , i.e., with $y, v \in \Gamma_x$ also $y + v \in \Gamma_x$ and $-y \in \Gamma_x$. It then follows (see Exercise 4) that Γ_x is a grid, generated by $k \leq d$ linearly independent vectors $g_1(x), \dots, g_k(x) \in \mathbb{R}^d$:

$$\Gamma_x = \{m_1 g_1(x) + \dots + m_k g_k(x); m_i \in \mathbb{Z}\}.$$

We extend $g_1(x), \dots, g_k(x)$ to a basis $g_1(x), \dots, g_d(x)$ of \mathbb{R}^d . Then, e induces a diffeomorphism

$$\begin{aligned} \mathbb{T}^k \times \mathbb{R}^{d-k} &\rightarrow M_x \\ (\theta_1, \dots, \theta_k, \tau_{k+1}, \dots, \tau_d) &\mapsto e\left(x, \sum_{i=1}^k \frac{\theta_i}{2\pi} g_i(x) + \sum_{j=k+1}^d \tau_j g_j(x)\right). \end{aligned}$$

If M_x is compact, then necessarily $k = d$ and M_x is a torus. The above map then becomes the bijection

$$\mathbb{T}^d \rightarrow M_x : \theta \mapsto e\left(x, \sum_{i=1}^d \frac{\theta_i}{2\pi} g_i(x)\right).$$

(b) Next we show that $g_i(x)$ is the gradient of some function $U_i(x)$. For notational convenience, we omit the subscript i and consider a differentiable function g with

$$e(x, g(x)) = e(x, 0), \quad x \in B,$$

or equivalently,

$$\ell \circ e(x, g(x)) = (x, 0), \quad x \in B.$$

Differentiating this relation gives (with I the d -dimensional identity)

$$A \begin{pmatrix} I \\ g'(x) \end{pmatrix} = \begin{pmatrix} I \\ 0 \end{pmatrix}$$

where A is the Jacobian matrix of $\ell \circ e$ at $(x, g(x))$. We thus have

$$(I \ g'(x)^T) A^T J A \begin{pmatrix} I \\ g'(x) \end{pmatrix} = (I \ 0) J \begin{pmatrix} I \\ 0 \end{pmatrix} = 0.$$

Since $\ell \circ e$ is a symplectic transformation, we have $A^T J A = J$, and hence the above equation reduces to

$$g'(x)^T - g'(x) = 0.$$

By the Integrability Lemma VI.2.7, there is a function U such that $g(x) = \nabla U(x)$. We may assume $U(x_0) = 0$.

(c) The result of (b) allows us to extend the bijection of (a) to a symplectic transformation. For this, we consider the generating function

$$S(x, \theta) = \sum_{i=1}^d \frac{\theta_i}{2\pi} U_i(x).$$

With $u(x) = (U_1(x), \dots, U_d(x))$, the mixed second derivative of S is

$$S_{x\theta}(x, \theta) = \frac{1}{2\pi} u_x(x) = \frac{1}{2\pi} (g_1(x), \dots, g_d(x)),$$

which is invertible because of the linear independence of the g_i . The equations

$$a = \frac{\partial S}{\partial \theta} = \frac{1}{2\pi} u(x), \quad y = \frac{\partial S}{\partial x} = \sum_{i=1}^d \frac{\theta_i}{2\pi} g_i(x)$$

define a bijective symplectic transformation (for some neighbourhood D of 0, and possibly with a reduced neighbourhood B of x_0)

$$\beta : D \times \mathbb{R}^d \rightarrow B \times \mathbb{R}^d : (a, \theta) \mapsto (x, y) = \left(f(a), \sum_{i=1}^d \frac{\theta_i}{2\pi} g_i(f(a)) \right)$$

where $x = f(a)$ is the inverse map of $a = \frac{1}{2\pi} u(x)$. We now define

$$\hat{\psi} = e \circ \beta : D \times \mathbb{R}^d \rightarrow \bigcup_{x \in B} M_x.$$

By construction, this map is smooth and symplectic, and such that $f_i(a) = x_i = F_i(p, q)$ for $(p, q) = \hat{\psi}(a, \theta)$. It is surjective by Theorem 1.4. By part (a) of this proof, it becomes injective when the θ_i are taken mod 2π , thus yielding a transformation ψ defined on $D \times \mathbb{T}^d$ with the stated properties. \square

X.1.4 Conditionally Periodic Flows

An immediate and important consequence of Theorem 1.6 is the following.

Corollary 1.9. *In the situation of Theorem 1.6, consider the completely integrable system with Hamiltonian $H = F_1$. In the action-angle variables (a, θ) , the Hamiltonian equations become*

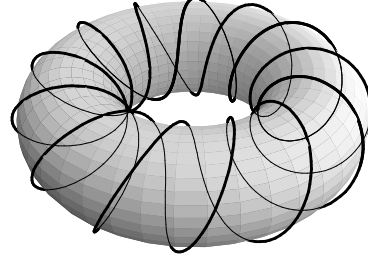
$$\dot{a}_i = 0, \quad \dot{\theta}_i = \omega_i(a) \quad (i = 1, \dots, d)$$

with $\omega_i(a) = \partial K / \partial a_i(a)$, where $K(a) = H(p, q)$ for $(p, q) = \psi(a, \theta)$.

The flow of a differential system

$$\dot{\theta} = \omega, \quad \omega = (\omega_i) \in \mathbb{R}^d$$

on the torus \mathbb{T}^d is called *conditionally periodic* with *frequencies* ω_i . The flow is periodic if there exist integers k_i such that for any two frequencies the relation $\omega_i/\omega_j = k_i/k_j$ holds. Otherwise, the flow is called *quasi-periodic*. In particular, the latter occurs when the frequencies are rationally independent, or *non-resonant*: the only integers k_i with $k_1\omega_1 + \dots + k_d\omega_d = 0$ are $k_1 = \dots = k_d = 0$. For non-resonant frequencies, it is well known (see Arnold (1989), p. 287) that every trajectory $\{\theta(t) : t \in \mathbb{R}\}$ is dense on the torus \mathbb{T}^d and uniformly distributed.



Example 1.10. We take up again the example of motion in a central field, Example 1.2. For given H and L , we now assume that

$$\{r > 0; 2(H - V(r)) - L^2/r^2 > 0\} = [r_0, r_1]$$

is a non-empty interval and the derivatives of $2(H - V(r)) - L^2/r^2$ are non-vanishing at r_0, r_1 . By (1.7), the motion from r_0 to r_1 and back again takes a time T and runs through an angle Φ which are given by

$$T = 2 \int_{r_0}^{r_1} \frac{1}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}} d\rho, \quad (1.12)$$

$$\Phi = 2 \int_{r_0}^{r_1} \frac{L/\rho^2}{\sqrt{2(H - V(\rho)) - L^2/\rho^2}} d\rho. \quad (1.13)$$

Note that r_0, r_1, T, Φ are functions of H and L . The solution is periodic if Φ is a rational multiple of 2π . This occurs for the Kepler problem, where $\Phi = 2\pi$ and where $T = 2\pi/(-2H)^{3/2}$ (for $H < 0$) depends only on H ; see Exercise I.5.

We now construct action-angle variables and compute the frequencies of the system. We begin by constructing the mapping $e(x, y)$ as defined by (1.11) for the variables $x = (x_1, x_2) = (H, L)$ and $y = (y_1, y_2)$ of (1.6). For a given (x, y) , we consider $(x, 0)$ and we fix (p, q) with $p = (p_r, p_\varphi)$ and $q = (r, \varphi)$ such that $\ell(p, q) = (x, 0)$, e.g., by choosing $r = r_0, \varphi = 0, p_r = 0, p_\varphi = L$. The mapping $e(x, y)$ is defined by the flow at time $t = 1$ corresponding to the Hamiltonian

$$F_y = y_1 H + y_2 L = y_1 \left(\frac{1}{2}(p_r^2 + p_\varphi^2/r^2) + V(r) \right) + y_2 p_\varphi,$$

i.e., by the solution at $t = 1$ of

$$\begin{aligned} \dot{p}_r &= -y_1 \frac{p_\varphi^2}{r^3} - y_1 V'(r), & \dot{p}_\varphi &= 0 \\ \dot{r} &= y_1 p_r, & \dot{\varphi} &= y_1 \frac{p_\varphi}{r^2} + y_2. \end{aligned} \quad (1.14)$$

If we denote the flow of the original system with Hamiltonian $H(p_r, p_\varphi, r, \varphi)$ by φ_t , then we have

$$e(x, y) = \varphi_{y_1}(0, L, r_0, 0) + (0, 0, 0, y_2)^T$$

with the last component taken modulo 2π . Hence, the values of y satisfying $e(x, y) = e(x, 0)$ are

$$y = m_1 g_1(x) + m_2 g_2(x)$$

with integers m_1, m_2 and

$$g_1 = \begin{pmatrix} T \\ -\Phi \end{pmatrix}, \quad g_2 = \begin{pmatrix} 0 \\ 2\pi \end{pmatrix}.$$

We know from the proof of Theorem 1.6 that g_1 and g_2 are the gradients of functions $U_1(H, L)$ and $U_2(H, L)$, respectively. Clearly, $U_2 = 2\pi L$. The expression for U_1 is less explicit. With the construction of the Integrability Lemma VI.2.7, this function is obtained by quadrature, in a neighbourhood of (H_0, L_0) , as

$$U_1(H, L) = \int_0^1 \left((H - H_0) T(H_0 + s(H - H_0), L_0 + s(L - L_0)) - (L - L_0) \Phi(H_0 + s(H - H_0), L_0 + s(L - L_0)) \right) ds.$$

(For the Kepler problem, $T = 2\pi/(-2H)^{3/2}$, $\Phi = 0 \bmod 2\pi$, and hence $U_1 = 2\pi/\sqrt{-2H}$.) For the action variables we thus obtain

$$a_1 = \frac{1}{2\pi} U_1(H, L), \quad a_2 = L.$$

The angle variables are given by $y = \frac{1}{2\pi}(\theta_1 g_1 + \theta_2 g_2)$, i.e.,

$$\theta_1 = y_1 \frac{2\pi}{T}, \quad \theta_2 = y_2 + y_1 \frac{\Phi}{T}. \quad (1.15)$$

Writing the total energy $H = K(a_1, L)$ if a_1 is given by the above formula, we obtain, by differentiation of the identity $2\pi a_1 = U_1(K(a_1, L), L)$,

$$2\pi = \frac{\partial U_1}{\partial H} \frac{\partial K}{\partial a_1}, \quad 0 = \frac{\partial U_1}{\partial H} \frac{\partial K}{\partial a_2} + \frac{\partial U_1}{\partial L}$$

and hence the frequencies

$$\omega_1 = \frac{\partial K}{\partial a_1} = \frac{2\pi}{T}, \quad \omega_2 = \frac{\partial K}{\partial a_2} = \frac{\Phi}{T}. \quad (1.16)$$

X.1.5 The Toda Lattice – an Integrable System

Our method is based on the realization that the Toda lattice belongs to a class of evolution equations which can be studied, and in some cases solved, by utilization of a certain associated eigenvalue problem.

(H. Flaschka 1974)

Classical examples of integrable systems from mechanics include Kepler's problem (Newton 1687/1713, Joh. Bernoulli 1710), the planar motion of a point mass attracted by two fixed centres (Euler 1760), Kepler's problem in a homogeneous force field (Lagrange 1766 solved this as the limit of the previous problem when one centre is at infinity), various spinning tops (Euler 1758b, Lagrange 1788, Kovalevskaya 1889, Goryachev 1899 and Chaplygin 1901), a number of integrable cases of the motion of a rigid body in a fluid, the motion of point vortices in the plane. We refer to Arnold, Kozlov & Neishtadt (1997) and Kozlov (1983) for interesting accounts of these problems and for further references.

Here we consider the celebrated example of the Toda lattice which was the starting point for a huge amount of work on integrable systems in the last few decades, with fascinating relationships to soliton theory in partial differential equations (most notably the Korteweg-de Vries equation) and to eigenvalue algorithms of Numerical Analysis; see Deift (1996) for an account of these developments.

The Toda lattice (or chain) is a system of particles on a line interacting pairwise with exponential forces. Such systems were studied by Toda (1970) as discrete models for nonlinear wave propagation. The motion is determined by the Hamiltonian

$$H(p, q) = \sum_{k=1}^n \left(\frac{1}{2} p_k^2 + \exp(q_k - q_{k+1}) \right). \quad (1.17)$$

Two types of boundary conditions have found particular attention in the literature:

(i) periodic boundary conditions: $q_{n+1} = q_1$;

(ii) put formally $q_{n+1} = +\infty$, so that the term $\exp(q_n - q_{n+1})$ does not appear. It was found by Hénon, Flaschka and independently Manakov in 1974 that the periodic Toda system is integrable. Moser (1975) then gave a detailed study of the non-periodic case (ii).

Flaschka (1974) introduced new variables

$$a_k = -\frac{1}{2} p_k, \quad b_k = \frac{1}{2} \exp\left(\frac{1}{2}(q_k - q_{k+1})\right).$$

(Take $b_n = 0$ in case (ii)). Along a solution $(p(t), q(t))$ of the Toda system, the corresponding functions $(a(t), b(t))$ satisfy the differential equations

$$\dot{a}_k = 2(b_k^2 - b_{k-1}^2), \quad \dot{b}_k = b_k(a_{k+1} - a_k)$$

(with $a_{n+1} = a_1$ in case (i), $b_n = 0$ in case (ii)). With the matrices

$$L = \begin{pmatrix} a_1 & b_1 & & & & b_n \\ b_1 & a_2 & b_2 & & 0 & \\ & b_2 & a_3 & b_3 & & \\ & & \ddots & \ddots & \ddots & \\ 0 & & & b_{n-2} & a_{n-1} & b_{n-1} \\ b_n & & & & b_{n-1} & a_n \end{pmatrix},$$

$$B = B(L) = \begin{pmatrix} 0 & b_1 & & & & -b_n \\ -b_1 & 0 & b_2 & & 0 & \\ & -b_2 & 0 & b_3 & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & -b_{n-2} & 0 & b_{n-1} \\ b_n & & & & -b_{n-1} & 0 \end{pmatrix},$$

the differential equations can be written in the *Lax pair* form

$$\dot{L} = BL - LB. \quad (1.18)$$

This system has an *isospectral flow*, that is, along any solution $L(t)$ of (1.18) the eigenvalues do not depend on t ; see Lemma IV.3.4. The eigenvalues $\lambda_1, \dots, \lambda_n$ of L are therefore first integrals of the Toda system. They are independent and turn out to be in involution, in a neighbourhood of every point where the λ_i are all different; see Exercise 6. Hence, the Toda lattice is a completely integrable system. Its Hamiltonian can be written as

$$H = \sum_{k=1}^n (2a_k^2 + 4b_k^2) = 2 \operatorname{trace} L^2 = 2 \sum_{i=1}^n \lambda_i^2.$$

We conclude this section with a numerical example for the periodic Toda lattice. We choose $n = 3$ and the initial conditions $p_1 = -1.5$, $p_2 = 1$, $p_3 = 0.5$ and $q_1 = 1$, $q_2 = 2$, $q_3 = -1$. We apply to the system with Hamiltonian (1.17) the symplectic second-order Störmer–Verlet method and the non-symplectic classical fourth-order Runge–Kutta method with two different step sizes. The left pictures of Fig. 1.3 show the numerical approximations to the eigenvalues, and the right pictures the deviations of the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ along the numerical solution from their initial values. Clearly, the eigenvalues are not invariants of the numerical schemes. However, Fig. 1.3 illustrates that the eigenvalues along the numerical solution remain close to their correct values over very long time intervals for the symplectic method, whereas they drift off for the non-symplectic method.

An explanation of the long-time near-preservation of the first integrals of completely integrable systems by symplectic methods will be given in the following sections, using backward error analysis and the perturbation theory for integrable Hamiltonian systems.

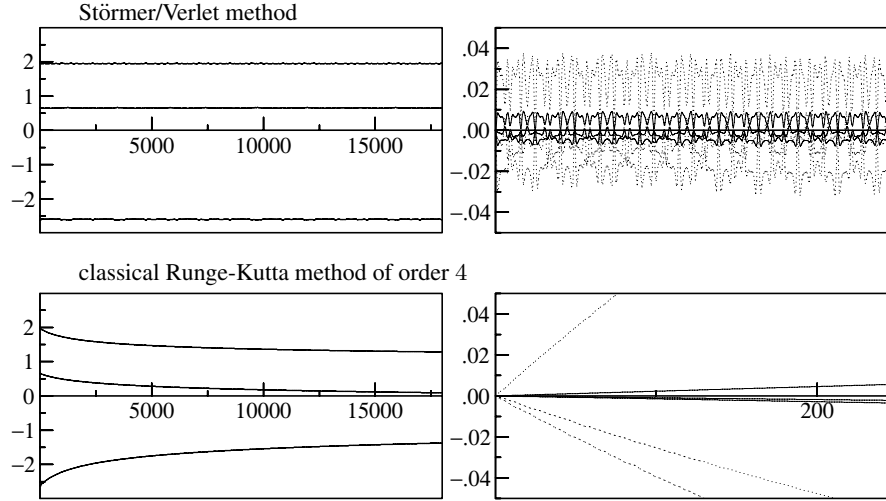


Fig. 1.3. Numerically obtained eigenvalues (left pictures) and errors in the eigenvalues (right pictures) for the step sizes $h = 0.1$ (dotted) and $h = 0.05$ (solid line)

X.2 Transformations in the Perturbation Theory for Integrable Systems

Problème général de la Dynamique. Nous sommes donc conduit à nous proposer le problème suivant: Étudier les équations canoniques

$$\frac{dx_i}{dt} = \frac{dF}{dy_i}, \quad \frac{dy_i}{dt} = -\frac{dF}{dx_i},$$

en supposant que la fonction F peut se développer suivant les puissances d'un paramètre très petit μ de la manière suivante:

$$F = F_0 + \mu F_1 + \mu^2 F_2 + \dots,$$

en supposant de plus que F_0 ne dépend que des x et est indépendant des y ; et que F_1, F_2, \dots sont des fonctions périodiques de période 2π par rapport aux y . (H. Poincaré 1892, p. 32f.)

Consider a small perturbation of a completely integrable Hamiltonian. In action-angle variables (a, θ) on $D \times \mathbb{T}^d$ (D an open subset of \mathbb{R}^d), this takes the form

$$H(a, \theta) = H_0(a) + \varepsilon H_1(a, \theta), \quad (2.1)$$

where ε is a small parameter. We assume that H_0 and H_1 are real-analytic, and that the perturbation H_1 (which may depend also on ε) is bounded by a constant on a complex neighbourhood of $D \times \mathbb{T}^d$ that is independent of ε . No other restriction shall be imposed on the perturbation.

For the unperturbed system ($\varepsilon = 0$) we have seen that the motion is conditionally periodic on invariant tori $\{a = \text{const.}, \theta \in \mathbb{T}^d\}$. Perturbation theory aims at an understanding of the flow of the perturbed system. The basic tools are symplectic

coordinate transformations which take the system to a form that allows the long-time behaviour (perpetually, or over time scales large compared to ε^{-1}) of solutions of the system (certain solutions, or all solutions with initial values in some ball) to be read off. There are different transformations that provide answers to these problems. The emphasis in this section will be on the construction of suitable transformations, not on the technical but equally important aspects of obtaining estimates for them.

The methods in Poincaré's *Méthodes Nouvelles* form the now classical part of perturbation theory, but the theories of Birkhoff, Siegel, Kolmogorov/Arnold/Moser (KAM) and Nekhoroshev in the 20th century have become "classics" in their own right.

X.2.1 The Basic Scheme of Classical Perturbation Theory

In the spirit of the preceding section, one might search for a symplectic change of coordinates $(a, \theta) \mapsto (b, \varphi)$ close to the identity such that the perturbed Hamiltonian written in the new variables (b, φ) depends only on b , or more modestly, depends only on b up to a remainder term of order $\mathcal{O}(\varepsilon^N)$ with a large $N > 1$, or to begin even more modestly, with $N = 2$. We search for a generating function

$$S(b, \theta) = b \cdot \theta + \varepsilon S_1(b, \theta)$$

where \cdot symbolizes the Euclidean product of vectors in \mathbb{R}^d and S_1 is 2π -periodic in θ . Naively, we require that the symplectic transformation defined by

$$a = \frac{\partial S}{\partial \theta}(b, \theta), \quad \varphi = \frac{\partial S}{\partial b}(b, \theta)$$

be such that the order- ε term in the expansion of the Hamiltonian in the new variables, $K(b, \varphi) = H(a, \theta)$, $K(b, \varphi) = H_0(b) + \varepsilon K_1(b, \varphi) + \dots$ depends only on b . Since

$$H(a, \theta) = H\left(b + \varepsilon \frac{\partial S_1}{\partial \theta}(b, \theta), \theta\right) = H_0(b) + \varepsilon \left\{ \omega(b) \cdot \frac{\partial S_1}{\partial \theta}(b, \theta) + H_1(b, \theta) \right\} + \dots$$

with the vector of frequencies

$$\omega(b) = \frac{\partial H_0}{\partial b}(b),$$

the function S_1 must satisfy the partial differential equation

$$\omega(b) \cdot \frac{\partial S_1}{\partial \theta}(b, \theta) + H_1(b, \theta) = \overline{H}_1(b) \quad (2.2)$$

for a function \overline{H}_1 that does not depend on θ . Since S_1 is required to be 2π -periodic in θ , the function \overline{H}_1 must equal the average of H_1 over the angles:

$$\overline{H}_1(b) = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} H_1(b, \theta) d\theta.$$

Equation (2.2) is the basic equation of Hamiltonian perturbation theory. From the Fourier series of S_1 and H_1 ,

$$S_1(b, \theta) = \sum_{k \in \mathbb{Z}^d} s_k(b) e^{ik \cdot \theta}, \quad H_1(b, \theta) = \sum_{k \in \mathbb{Z}^d} h_k(b) e^{ik \cdot \theta}$$

we obtain a formal solution of (2.2) by comparing Fourier coefficients: $s_0(b)$ is arbitrary and

$$s_k(b) = -\frac{h_k(b)}{ik \cdot \omega(b)}, \quad k \neq 0. \quad (2.3)$$

At this point, however, we are struck by the *problem of small denominators*. For any values of the frequencies $\omega_j(b)$, the denominator $k \cdot \omega(b) = k_1 \omega_1(b) + \dots + k_d \omega_d(b)$ becomes arbitrarily small for some $k = (k_1, \dots, k_d) \in \mathbb{Z}^d$, and even vanishes if the frequencies are rationally dependent.

For a perturbation where only finitely many Fourier coefficients h_k are non-zero, the construction above excludes only a finite number of resonant frequencies (i.e., those with $k \cdot \omega(b) = 0$ for a $k \in \mathbb{Z}^d$ with $h_k \neq 0$) and small neighbourhoods around them. For $\omega(b)$ outside these neighbourhoods and for φ on a complex neighbourhood of \mathbb{T}^d , we obtain for the Hamiltonian in the new variables

$$K(b, \varphi) = H_0(b) + \varepsilon \overline{H}_1(b) + \mathcal{O}(\varepsilon^2).$$

In the general case, we can approximate the perturbation H_1 up to $\mathcal{O}(\varepsilon^2)$ by a trigonometric polynomial. For analytic H_1 , the Fourier coefficients h_k decay exponentially with $|k| = \sum_i |k_i|$, and hence the required degree m of the approximating trigonometric polynomial grows logarithmically with ε , i.e., $m \sim |\log \varepsilon|$.

As $\varepsilon \rightarrow 0$, the remainder term is under control only for those frequencies $\omega = \omega(b)$ for which the exponentially decaying Fourier coefficients h_k of the perturbation decay faster than the denominators $ik \cdot \omega$ with growing $|k|$. This is certainly the case for frequencies satisfying *Siegel's diophantine condition* (or *strong non-resonance condition*, as it is sometimes called)

$$|k \cdot \omega| \geq \gamma |k|^{-\nu}, \quad k \in \mathbb{Z}^d, k \neq 0 \quad (2.4)$$

for some positive constants γ, ν . (Here again, $|k| = \sum_i |k_i|$). If $\nu > d - 1$, the set of frequencies in a fixed ball that do *not* satisfy (2.4) has Lebesgue measure bounded by $\text{Const} \cdot \gamma$ (Exercise 5). Therefore, almost all frequencies satisfy (2.4) for some $\gamma > 0$. However, for any γ and ν , the complementary set is open and dense in \mathbb{R}^d .

X.2.2 Lindstedt–Poincaré Series

... pour que la méthode de M. Lindstedt soit applicable, soit sous sa forme primitive, soit sous celle que je lui ai ensuite donnée, il faut qu'en première approximation les moyens mouvements ne soient liés par aucune relation linéaire à coefficients entiers; ...

Il semble donc permis de conclure que les séries (...) ne convergent pas. Toutefois le raisonnement qui précède ne suffit pas pour établir ce point avec une rigueur complète. (H. Poincaré 1893, pp. vi, 103.)



Fig. 2.1. Henri Poincaré (left), born: 29 April 1854 in Nancy (France), died: 17 July 1912 in Paris; Anders Lindstedt (right), born: 27 June 1854 in Sundborn (Sweden), died: 1939. Reproduced with permission of Bibl. Math. Univ. Genève

The above construction is extended without any additional difficulty to arbitrary finite order in ε . The generating function is now sought in the form

$$S(b, \theta) = b \cdot \theta + \varepsilon S_1(b, \theta) + \varepsilon^2 S_2(b, \theta) + \dots + \varepsilon^{N-1} S_{N-1}(b, \theta) \quad (2.5)$$

and, as before, the requirement that the first N terms in the ε -expansion of the Hamiltonian in the new variables be independent of the angles, leads via a Taylor expansion of the Hamiltonian to equations of the form (2.2) for S_1, \dots, S_{N-1} :

$$\omega(b) \cdot \frac{\partial S_j}{\partial \theta} + K_j(b, \theta) = \overline{K}_j(b) \quad (2.6)$$

where $K_1 = H_1$,

$$K_2 = \frac{1}{2} \frac{\partial^2 H_0}{\partial a^2} \left(\frac{\partial S_1}{\partial \theta}, \frac{\partial S_1}{\partial \theta} \right) + \frac{\partial H_1}{\partial a} \cdot \frac{\partial S_1}{\partial \theta},$$

and in general, K_j is a sum of terms

$$\frac{1}{i!} \frac{\partial^i H_{k_0}}{\partial a^i} \left(\frac{\partial S_{k_1}}{\partial \theta}, \dots, \frac{\partial S_{k_i}}{\partial \theta} \right) \quad \text{with } k_0 + k_1 + \dots + k_i = j.$$

The function \overline{K}_j denotes again the angular average of K_j . These equations can be formally solved in the case of rationally independent frequencies. The Hamiltonian in the new variables is then

$$K(b, \varphi) = H_0(b) + \varepsilon \overline{K}_1(b) + \varepsilon^2 \overline{K}_2(b) + \dots + \varepsilon^{N-1} \overline{K}_{N-1}(b) + \varepsilon^N R_N(b, \theta). \quad (2.7)$$

The possible convergence of the series for $N \rightarrow \infty$ is a delicate issue that was not resolved conclusively by Poincaré (1893) in his chapter on “Divergence des séries de M. Lindstedt”. If for some b^* , the series (2.5) together with its partial derivatives converged as $N \rightarrow \infty$, then $\{b = b^*, \varphi \in \mathbb{T}^d\}$ would be an invariant torus of the perturbed Hamiltonian system. However, it was not until Kolmogorov (1954) that the existence of invariant tori – for diophantine frequencies – was found, using a different construction. A direct proof of the convergence of the series of classical perturbation theory for diophantine frequencies was obtained only in 1988 by Eliasson (published in 1996); also see Giorgilli & Locatelli (1997) and references therein.

Nevertheless, already the truncated series (2.5) leads in a rather simple way to strong conclusions about the flow over long time scales when it is combined with the idea of approximating the Hamiltonian by a trigonometric polynomial: the “ultra-violet cut-off”, an idea briefly addressed by Poincaré (1893), p. 98f., and taken to its full bearing by Arnold (1963) in his proof of the KAM theorem. We formulate a lemma for a fixed truncation index N . Here, $\omega_{\varepsilon, N}(b)$ denotes the derivative of the truncated series (2.7) with respect to b .

Lemma 2.1. *Suppose that $\omega(b^*)$ satisfies the diophantine condition (2.4). For any fixed $N \geq 2$, there are positive constants ε_0, c, C such that the following holds for $\varepsilon \leq \varepsilon_0$: there exists a real-analytic symplectic change of coordinates $(a, \theta) \mapsto (b, \varphi)$ such that every solution $(b(t), \varphi(t))$ of the perturbed system in the new coordinates, starting with $\|b(0) - b^*\| \leq c |\log \varepsilon|^{-\nu-1}$, satisfies*

$$\begin{aligned} \|b(t) - b(0)\| &\leq C t \varepsilon^N \quad \text{for } t \leq \varepsilon^{-N+1}, \\ \|\varphi(t) - \omega_{\varepsilon, N}(b(0))t - \varphi(0)\| &\leq C (t^2 + t |\log \varepsilon|^{\nu+1}) \varepsilon^N \quad \text{for } t^2 \leq \varepsilon^{-N+1}. \end{aligned}$$

Moreover, the transformation is $\mathcal{O}(\varepsilon)$ -close to the identity: $\|(a, \theta) - (b, \varphi)\| \leq C\varepsilon$ holds for (a, θ) and (b, φ) related by the above coordinate transform, for $\|b - b^*\| \leq c |\log \varepsilon|^{-\nu-1}$ and for φ in an ε -independent complex neighbourhood of \mathbb{T}^d .

The constants ε_0, c, C depend on N, d, γ, ν and on bounds of H_0 and H_1 on a complex neighbourhood of $\{b^*\} \times \mathbb{T}^d$.

Proof. Using the relations (2.3) and their analogues for (2.6), it is a straightforward but somewhat tedious exercise to show that at the given particular b^* , the functions $K_j(b^*, \cdot), S_j(b^*, \cdot)$ are all analytic on the same complex neighbourhood of \mathbb{T}^d , and that the remainder term is bounded by

$$|R_N(b^*, \theta)| \leq C = C(N, d, \gamma, \nu)$$

for all θ in a complex neighbourhood of \mathbb{T}^d which is independent of ε . Here, C depends in addition on the bound of H_1 on a complex neighbourhood of $\{b^*\} \times \mathbb{T}^d$, or what amounts to the same by Cauchy’s estimates, on bounds of the exponential decay of the Fourier coefficients h_k of H_1 . (In case of doubt, see also Sect. X.4 for explicit estimates.)

Assume first that $H_1(b, \theta)$ is a trigonometric polynomial in θ of degree m . Then K_j, S_j are trigonometric polynomials of degree jm . Since $|k \cdot \omega(b)| \geq |k \cdot \omega(b^*)| - |k|(\max \|\omega'\|)\|b - b^*\|$, there is a $\delta > 0$ such that

$$|k \cdot \omega(b)| \geq \frac{1}{2}\gamma |k|^{-\nu} \quad \text{for } \|b - b^*\| \leq \delta, \quad |k| \leq Nm.$$

This number δ is proportional to $\gamma(Nm)^{-\nu-1}$. Consequently, since the construction involves only the trigonometric polynomials K_j, S_j of degree up to Nm , the above estimate for the remainder term R_N holds also for $\|b - b^*\| \leq \delta$. To approximate a general analytic H_1 by trigonometric polynomials up to $\mathcal{O}(\varepsilon^N)$, we must choose the degree m proportional to $|\log \varepsilon^N|$. With the choice $\delta = c(N^2 |\log \varepsilon|)^{-\nu-1}$, for a sufficiently small $c > 0$ independent of ε (and N), the above bound for the remainder $R_N(b, \theta)$ is then valid for b in the complex ball $\|b - b^*\| \leq 2\delta$ and for φ in a complex neighbourhood of \mathbb{T}^d (which depends only on N). By Cauchy's estimates, this implies

$$\left\| \frac{\partial R_N}{\partial \theta}(b, \theta) \right\| \leq C, \quad \left\| \frac{\partial R_N}{\partial b}(b, \theta) \right\| \leq \frac{C}{\delta}$$

for $\|b - b^*\| \leq \delta$ and $\theta \in \mathbb{T}^d$. Hence, as long as $\|b(t) - b^*\| \leq \delta$, the Hamiltonian differential equations are of the form

$$\dot{b} = -\frac{\partial K}{\partial \varphi} = -\varepsilon^N \frac{\partial R_N}{\partial \theta} \frac{\partial \theta}{\partial \varphi} = \mathcal{O}(\varepsilon^N), \quad \dot{\varphi} = \frac{\partial K}{\partial b} = \omega_{\varepsilon, N}(b) + \mathcal{O}(\varepsilon^N/\delta).$$

This implies the result. \square

Hence, the tori $\{b = b(0), \varphi \in \mathbb{T}^d\}$ are nearly invariant over a time scale ε^{-N+1} , and the flow is close to a quasiperiodic flow over times bounded by the square root of ε^{-N+1} . Lemma 2.1 is just a preliminary to more substantial results (which hold under appropriate additional conditions): invariant tori carrying a quasiperiodic flow with diophantine frequencies persist under small Hamiltonian perturbations (Kolmogorov 1954); every solution of the perturbed system remains close, within a positive power of ε , to some torus over times that are exponentially long in a negative power of ε (Nekhoroshev 1977); solutions starting close to an invariant torus with diophantine frequencies stay within twice the initial distance over time intervals that are exponentially long in a negative power of the distance (Perry & Wiggins 1994) or even exponentially long in the exponential of the inverse of the distance (Morbidielli & Giorgilli 1995).

The symplectic transformations of this subsection were constructed using the mixed-variable generating function $S(b, \theta)$. As was pointed out for example by Benettin, Galgani & Giorgilli (1985), rigorous estimates for the remainder terms are often obtained in a simpler way using the *Lie method*, which involves constructing the near-identity symplectic transformation as the time- ε flow of some auxiliary Hamiltonian system with a suitably defined Hamiltonian $\chi(b, \varphi)$. As before, the condition that the Hamiltonian $H(a, \theta) = K(b, \varphi)$ should depend on φ only in higher-order terms, leads to equations of the form (2.2), now for χ instead of S_1 . We will use such a construction in the following subsection.

X.2.3 Kolmogorov's Iteration

It is easy to grasp the meaning of Theorem 1 for mechanics. It indicates that an s -parametric family of conditionally periodic motions [...] cannot, under conditions (3) and (4) [here: (2.4) and (2.9)], disappear as a result of a small change in the Hamilton function H .

In this note we confine ourselves to the construction of the transformation. (A.N. Kolmogorov 1954)

For the completely integrable Hamiltonian $H_0(a)$, the phase space is foliated into invariant tori parametrized by a . We now fix one such torus $\{a = a^*, \theta \in \mathbb{T}^d\}$ with strongly diophantine frequencies $\omega = \omega(a^*)$. Without loss of generality, we may assume $a^* = 0$. This particular torus is invariant under the flow of every Hamiltonian $H(a, \theta)$ for which the linear terms in the Taylor expansion with respect to a at 0 are independent of θ :

$$H(a, \theta) = c + \omega \cdot a + \frac{1}{2} a^T M(a, \theta) a \quad (2.8)$$

with $c \in \mathbb{R}$, $\omega \in \mathbb{R}^d$, and a real symmetric $d \times d$ -matrix $M(a, \theta)$ analytic in its arguments. Since the Hamiltonian equations are of the form

$$\dot{a} = \mathcal{O}(\|a\|^2), \quad \dot{\theta} = \omega + \mathcal{O}(\|a\|),$$

the torus $\{a = 0, \theta \in \mathbb{T}^d\}$ is invariant and the flow on it is quasi-periodic with frequencies ω .

Consider now an analytic perturbation of such a Hamiltonian: $H(a, \theta) + \varepsilon G(a, \theta)$ with a small ε . Kolmogorov (1954) found a near-identity symplectic transformation $(a, \theta) \mapsto (\tilde{a}, \tilde{\theta})$, constructed by an iterative procedure, such that the perturbed Hamiltonian in the new variables is again of the form (2.8) with the same ω , and hence has the invariant torus $\{\tilde{a} = 0, \tilde{\theta} \in \mathbb{T}^d\}$ carrying a quasi-periodic flow with the frequencies of the unperturbed system. This holds under the conditions that ω satisfies the diophantine condition (2.4), and that the angular average

$$\overline{M}_0 := \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} M(0, \theta) d\theta \quad \text{is an invertible matrix.} \quad (2.9)$$

Here we describe the iterative construction of this symplectic transformation. The proof of convergence of the iteration will be given in Sect. X.5.

We construct a symplectic transformation $(a, \theta) \mapsto (b, \varphi)$ as the time- ε flow of an auxiliary Hamiltonian of the form

$$\chi(b, \varphi) = \xi \cdot \varphi + \chi_0(\varphi) + \sum_{i=1}^d b_i \chi_i(\varphi), \quad (2.10)$$

where $\xi \in \mathbb{R}^d$ is a constant vector, and $\chi_0, \chi_1, \dots, \chi_d$ are 2π -periodic functions. (Quadratic and higher-order terms in b play no role in the construction and are therefore omitted right at the outset.) The old and new coordinates are then related by

$$a = b + \varepsilon \frac{\partial \chi}{\partial \varphi}(b, \varphi) + \mathcal{O}(\varepsilon^2), \quad \theta = \varphi - \varepsilon \frac{\partial \chi}{\partial b}(b, \varphi) + \mathcal{O}(\varepsilon^2).$$

We insert this into

$$H(a, \theta) + \varepsilon G(a, \theta) = c + \omega \cdot b + \frac{1}{2} b^T M(b, \varphi) b + \varepsilon \left\{ \omega \cdot \frac{\partial \chi}{\partial \varphi}(b, \varphi) + b^T M(b, \varphi) \frac{\partial \chi}{\partial \varphi}(b, \varphi) + G(b, \varphi) \right\} + \mathcal{O}(\varepsilon \|b\|^2) + \mathcal{O}(\varepsilon^2).$$

We now require that the term in curly brackets be $Const + \mathcal{O}(\|b\|^2)$. Writing down the Taylor expansion

$$G(b, \varphi) = G_0(\varphi) + \sum_{i=1}^d b_i G_i(\varphi) + b^T Q(b, \varphi) b \quad (2.11)$$

and inserting the above ansatz for χ , this condition becomes

$$\begin{aligned} \omega \cdot \frac{\partial \chi_0}{\partial \varphi}(\varphi) + \sum_{i=1}^d b_i \left(\omega \cdot \frac{\partial \chi_i}{\partial \varphi}(\varphi) + u_i(\varphi) + v_i(\varphi) \right) \\ + G_0(\varphi) + \sum_{i=1}^d b_i G_i(\varphi) = Const., \end{aligned}$$

where $u = (u_1, \dots, u_d)^T$ and $v = (v_1, \dots, v_d)^T$ are defined by

$$u(\varphi) = M(0, \varphi) \xi, \quad (2.12)$$

$$v(\varphi) = M(0, \varphi) \frac{\partial \chi_0}{\partial \varphi}(\varphi). \quad (2.13)$$

The condition is fulfilled if

$$\omega \cdot \frac{\partial \chi_0}{\partial \varphi}(\varphi) + G_0(\varphi) = \overline{G}_0 \quad (2.14)$$

$$\omega \cdot \frac{\partial \chi_i}{\partial \varphi}(\varphi) + u_i(\varphi) + v_i(\varphi) + G_i(\varphi) = \overline{u}_i + \overline{v}_i + \overline{G}_i \quad (2.15)$$

$$\overline{u}_i + \overline{v}_i + \overline{G}_i = 0 \quad (i = 1, \dots, d). \quad (2.16)$$

Here the bars again denote angular averages. Note that equations (2.14), (2.15) are of the form (2.2). Equation (2.14) determines χ_0 and hence $v = (v_1, \dots, v_d)^T$ by (2.13). Equations (2.16) then give $\overline{u} = (\overline{u}_1, \dots, \overline{u}_d)^T$. By (2.12), we need

$$\overline{u} = \overline{M}_0 \xi,$$

which determines ξ uniquely because \overline{M}_0 is assumed to be invertible. Equation (2.12) then yields $u = (u_1, \dots, u_d)^T$. Finally, (2.15) determines χ_1, \dots, χ_d , and the construction of $\chi(b, \varphi)$ is complete. In the new variables (b, φ) , the perturbed Hamiltonian then takes the form

$$H(a, \theta) + \varepsilon G(a, \theta) = \widehat{c} + \omega \cdot b + \frac{1}{2} b^T \widehat{M}(b, \varphi) b + \varepsilon^2 \widehat{G}(b, \varphi) \quad (2.17)$$

with unchanged frequencies ω and with $\widehat{M}(b, \varphi) = M(b, \varphi) + \mathcal{O}(\varepsilon)$. The perturbation to the form (2.8) is thus reduced from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$. The iteration of this procedure turns out to be convergent, see Sect. X.5. This finally yields a symplectic change of coordinates that transforms the perturbed Hamiltonian to the form (2.8). The perturbed system thus has an invariant torus carrying a quasi-periodic flow with frequencies ω – a KAM torus, as it is named after Kolmogorov, Arnold and Moser.

X.2.4 Birkhoff Normalization Near an Invariant Torus

KAM tori are very sticky.
(A.D. Perry & S. Wiggins 1994)

In this subsection we describe a transformation studied by Pöschel (1993) and Perry & Wiggins (1994) for systems with Hamiltonian in the Kolmogorov form (2.8) in a neighbourhood of the invariant torus $\{a = 0, \theta \in \mathbb{T}^d\}$. This transformation is an analogue of a transformation of Birkhoff (1927) for Hamiltonian systems near an elliptic stationary point.

The symplectic change of coordinates $(a, \theta) \mapsto (b, \varphi)$ considered here transforms a Hamiltonian (2.8) with diophantine frequencies ω to the form $H(a, \theta) = K_N(b) + \mathcal{O}(\|b\|^N)$ for arbitrary N , or more precisely, the Hamiltonian in the new variables, $H_N(b, \varphi) = H(a, \theta)$, is of the form

$$H_N(b, \varphi) = \omega \cdot b + Z_N(b) + R_N(b, \varphi) \quad (2.18)$$

with $Z_N(b) = \mathcal{O}(\|b\|^2)$ and $R_N(b, \varphi) = \mathcal{O}(\|b\|^N)$. (We have taken the irrelevant constant term in (2.8) $c = 0$.) The equations of motion then take the form

$$\dot{b} = \mathcal{O}(\|b\|^N), \quad \dot{\varphi} = \omega + \mathcal{O}(\|b\|).$$

Therefore, in these variables $\{b = 0, \varphi \in \mathbb{T}^d\}$ is an invariant torus, and for sufficiently small r ,

$$\|b(0)\| \leq r \quad \text{implies} \quad \|b(t)\| \leq 2r \quad \text{for } t \leq C_N r^{-N+1}.$$

A judicious choice of N even yields time intervals that are exponentially long in a negative power of r on which solutions starting at a distance r stay within twice the initial distance (Perry & Wiggins 1994). Motion away from the torus can thus be only very slow.

The normal form (2.18) is constructed iteratively. Each iteration step is very similar to the procedure in Sect. X.2.1, where now the distance to the torus plays the role of the small parameter. Consider a Hamiltonian

$$H(a, \theta) = \omega \cdot a + Z(a) + R(a, \theta)$$

where $Z(a) = \mathcal{O}(\|a\|^2)$ and $R(a, \theta) = \mathcal{O}(\|a\|^k)$ for some $k \geq 2$ in a complex neighbourhood of $\{0\} \times \mathbb{T}^d$. We construct a symplectic change of coordinates $(a, \theta) \mapsto (b, \varphi)$ via a generating function $b \cdot \theta + S(b, \theta)$ as

$$a = b + \frac{\partial S}{\partial \theta}(b, \theta), \quad \varphi = \theta + \frac{\partial S}{\partial b}(b, \theta).$$

We expand (omitting the arguments (b, θ) in $\partial S/\partial \theta$ and $\partial H/\partial a$)

$$\begin{aligned} H\left(b + \frac{\partial S}{\partial \theta}, \theta\right) &= H(b, \theta) + \frac{\partial H}{\partial a} \cdot \frac{\partial S}{\partial \theta} + Q(b, \theta) \\ &= \omega \cdot b + Z(b) + \left\{ R(b, \theta) + \frac{\partial H}{\partial a} \cdot \frac{\partial S}{\partial \theta} \right\} + Q(b, \theta), \end{aligned}$$

where $|Q(b, \theta)| \leq \text{Const.} \|\partial S/\partial \theta\|^2$. Since $\partial H/\partial b = \omega + \mathcal{O}(\|b\|)$, we can make the expression in curly brackets independent of θ up to $\mathcal{O}(\|b\|^{k+1})$ by determining S from the equation of the form (2.2):

$$\omega \cdot \frac{\partial S}{\partial \theta}(b, \theta) + R(b, \theta) = \overline{R}(b).$$

For diophantine frequencies ω , we obtain $S(b, \theta) = \mathcal{O}(\|b\|^k)$ on a (reduced) complex neighbourhood of $\{0\} \times \mathbb{T}^d$ from the corresponding estimate for $R(b, \theta)$. It follows that the above symplectic transformation with generating function $b \cdot \theta + S(b, \theta)$ is well-defined for small $\|b\|$, and the Hamiltonian in the new variables, $\widehat{H}(b, \varphi) = H(a, \theta)$, becomes

$$\widehat{H}(b, \varphi) = \omega \cdot b + \widehat{Z}(b) + \widehat{R}(b, \varphi)$$

with $\widehat{Z}(b) = Z(b) + \overline{R}(b)$ and

$$\widehat{R}(b, \varphi) = \left(\frac{\partial H}{\partial a}(b, \theta) - \omega \right) \cdot \frac{\partial S}{\partial \theta}(b, \theta) + Q(b, \theta) = \mathcal{O}(\|b\|^{k+1}),$$

so that the order in b of the remainder term is augmented by 1. The procedure can be iterated, but unlike the iteration of the preceding subsection, this iteration is in general divergent. Nevertheless, a suitable finite termination yields remainder terms that are exponentially small in a positive power of r for $\|b\| \leq r$, by arguments similar to those of Sect. X.4.

X.3 Linear Error Growth and Near-Preservation of First Integrals

In the remaining part of this chapter we study the long-time behaviour of symplectic discretizations of integrable and near-integrable Hamiltonian systems. While here we will be concerned with general symplectic methods, it should be noted that some integrable problems admit integrable discretizations; see Suris (2003).

In this section we are concerned with the error growth of symplectic numerical methods and their approximate preservation of first integrals. A preliminary analysis of linear error growth for the Kepler problem was first given by Calvo & Sanz-Serna

(1993). Using backward error analysis and KAM theory, Calvo & Hairer (1995a) then showed linear error growth of symplectic methods applied to integrable systems when the frequencies at the initial value satisfy a diophantine condition (2.4). Here we give such a result under milder conditions on the initial values, combining backward error analysis and Lemma 2.1. We derive also a first result on the long-time near-preservation of all first integrals, which will be extended to exponentially long times in Sections X.4.3 and X.5.2 (under stronger assumptions on the starting values), and perpetually in Sect. X.6 (only for a Cantor set of step sizes).

Figure 3.1 illustrates the linear error growth of the symplectic Störmer–Verlet method, as opposed to the quadratic error growth for the classical fourth-order Runge–Kutta method, on the example of the Toda lattice. The same number of function evaluations was used for both methods.

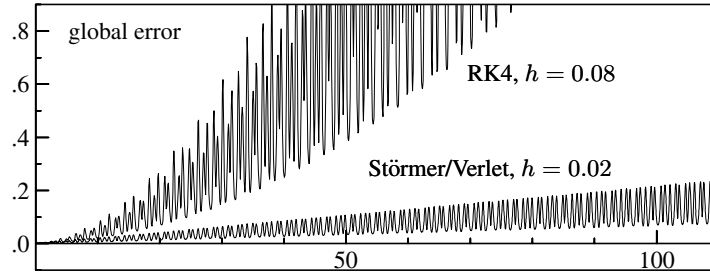


Fig. 3.1. Euclidean norm of the global error for the Störmer–Verlet scheme (step size $h = 0.02$) and the classical Runge–Kutta method of order 4 (step size $h = 0.08$) applied to the Toda lattice with $n = 3$ and initial values as in Fig. 1.3

We consider a completely integrable Hamiltonian system (usually not given in action-angle variables)

$$\dot{p} = -\frac{\partial H}{\partial q}(p, q), \quad \dot{q} = \frac{\partial H}{\partial p}(p, q) \quad (3.1)$$

and apply to it a symplectic numerical method with step size h , yielding a numerical solution sequence (p_n, q_n) . We assume that the Hamiltonian is real-analytic and that the conditions of the Arnold–Liouville theorem, Theorem 1.6, are fulfilled. Consider the symplectic transformation $(p, q) = \psi(a, \theta)$ to action-angle variables. We denote the inverse transformation as

$$(a, \theta) = (I(p, q), \Theta(p, q)). \quad (3.2)$$

We recall that the components I_1, \dots, I_d of $I = (I_i)$ are first integrals of the system: $I(p(t), q(t)) = I(p_0, q_0)$ for all t . In the action-angle variables, the Hamiltonian is $\mathcal{H}(a) = H(p, q)$, and we denote the frequencies

$$\omega(a) = \frac{\partial \mathcal{H}}{\partial a}(a). \quad (3.3)$$

We consider this in a neighbourhood of some $a^* \in \mathbb{R}^d$.

Theorem 3.1. *Consider applying a symplectic numerical integrator of order p to the completely integrable Hamiltonian system (3.1). Suppose that $\omega(a^*)$ satisfies the diophantine condition (2.4). Then, there exist positive constants C, c and h_0 such that the following holds for all step sizes $h \leq h_0$: every numerical solution starting with $\|I(p_0, q_0) - a^*\| \leq c|\log h|^{-\nu-1}$ satisfies*

$$\begin{aligned} \|(p_n, q_n) - (p(t), q(t))\| &\leq C t h^p \\ \|I(p_n, q_n) - I(p_0, q_0)\| &\leq C h^p \end{aligned} \quad \text{for } t = nh \leq h^{-p}.$$

The constants h_0, c, C depend on d, γ, ν , on bounds of the real-analytic Hamiltonian H on a complex neighbourhood of the torus $\{(p, q); I(p, q) = a^*\}$, and on the numerical method.

Proof. (a) In the action-angle variables (a, θ) , the exact flow is given as

$$a(t) = a(0), \quad \theta(t) = \omega(a(0))t + \theta(0). \quad (3.4)$$

By Theorem IX.3.1 (and Theorem IX.1.2), the truncated modified equation of the numerical method is Hamiltonian with¹

$$\tilde{H}(p, q) = H(p, q) + h^p H_{p+1}(p, q) + \dots + h^r H_{r+1}(p, q).$$

We choose $r = 2p$, and we denote by $(\tilde{p}(t), \tilde{q}(t))$ the solution of the modified equations with initial values (p_0, q_0) . In the variables (a, θ) , the modified Hamiltonian becomes $\tilde{H}(p, q) = \tilde{\mathcal{H}}(a, \theta)$ with

$$\tilde{\mathcal{H}}(a, \theta) = \mathcal{H}(a) + \varepsilon \mathcal{G}_h(a, \theta), \quad (3.5)$$

where $\varepsilon = h^p$ and the perturbation function \mathcal{G}_h is bounded independently of h on a complex neighbourhood of $\{a^*\} \times \mathbb{T}^d$. By Lemma 2.1 with $\varepsilon = h^p$ and $N \geq 3$, there is a symplectic change of coordinates $\mathcal{O}(h^p)$ -close to the identity, such that the solution of the modified equation in the new variables (b, φ) is of the form

$$\begin{aligned} \tilde{b}(t) &= \tilde{b}(0) + \mathcal{O}(th^{pN}), \\ \tilde{\varphi}(t) &= \omega_h(\tilde{b}(0))t + \tilde{\varphi}(0) + \mathcal{O}(th^{pN-1} + t^2h^{pN}) \end{aligned} \quad \text{for } t \leq h^{-p}, \quad (3.6)$$

with $\omega_h(b) = \omega(b) + \mathcal{O}(h^p)$. The constants symbolized by the \mathcal{O} -notation are independent of h , of $t \leq h^{-p}$ and of $(\tilde{b}(0), \tilde{\varphi}(0))$ with $|\tilde{b}(0) - a^*| \leq c|\log h|^{-\nu-1}$. Since the transformation between the variables (a, θ) and (b, φ) is $\mathcal{O}(h^p)$ close to the identity, it follows that the flow of the modified equations in the variables (a, θ) satisfies

¹ We always assume, without further mention, that the modified Hamiltonian is well-defined on the same open set D as the original Hamiltonian. This is true for arbitrary symplectic methods if D is simply connected; on general domains it is satisfied for (partitioned) Runge–Kutta methods and for splitting methods; see Sections IX.3 and IX.4.

$$\begin{aligned}\tilde{a}(t) &= \tilde{a}(0) + \mathcal{O}(h^p), \\ \tilde{\theta}(t) &= \omega(\tilde{a}(0))t + \tilde{\theta}(0) + te_h + \mathcal{O}(h^p)\end{aligned}\quad \text{for } 1 \leq t \leq h^{-p},$$

where $e_h = \omega_h(\tilde{b}(0)) - \omega(\tilde{a}(0)) = \mathcal{O}(h^p)$ yields the dominant contribution to the error. By comparison with (3.4) and since $\tilde{a}(t) = I(\tilde{p}(t), \tilde{q}(t))$, the difference between the exact solution and the solution of the modified equation therefore satisfies

$$\begin{aligned}(\tilde{p}(t), \tilde{q}(t)) - (p(t), q(t)) &= \mathcal{O}(th^p) \\ I(\tilde{p}(t), \tilde{q}(t)) - I(p_0, q_0) &= \mathcal{O}(h^p)\end{aligned}\quad \text{for } 1 \leq t \leq h^{-p}.$$

The same bounds for $t \leq 1$ follow by standard error estimates.

(b) It remains to bound the difference between the solution of the modified equation and the numerical solution. By construction of the modified equation with $r = 2p$ and by comparison with (3.6), one step of the method is of the form

$$b_{n+1} = b_n + \mathcal{O}(h^{r+1}), \quad \varphi_{n+1} = \omega_h(b_n)h + \varphi_n + \mathcal{O}(h^{r+1}).$$

It follows that for $t = nh$,

$$b_n = \tilde{b}(t) + \mathcal{O}(th^r), \quad \varphi_n = \tilde{\varphi}(t) + \mathcal{O}(t^2h^r).$$

For $t \leq h^{-p}$ and $r = 2p$, we have $th^r \leq h^p$. Hence the difference between the numerical solution and the solution of the modified equations in the original variables (p, q) is bounded by

$$\begin{aligned}(p_n, q_n) - (\tilde{p}(t), \tilde{q}(t)) &= \mathcal{O}(th^p) \\ I(p_n, q_n) - I(\tilde{p}(t), \tilde{q}(t)) &= \mathcal{O}(h^p)\end{aligned}\quad \text{for } t = nh \leq h^{-p}.$$

Together with the bound of part (a) this gives the result. \square

Remark 3.2. The linear error growth holds also when the symplectic method is applied to a perturbed integrable system with a perturbation parameter ε bounded by a positive power of the step size: $\varepsilon \leq K h^\alpha$ for some $\alpha > 0$. The proof of this generalization is the same as above, except that possibly a larger N is required in using Lemma 2.1.

Example 3.3 (Linear Error Growth for the Kepler Problem). From Example 1.10 we know that for the Kepler problem the frequencies (1.16) do not satisfy the diophantine condition (2.4). Nevertheless we observed a linear error growth for symplectic methods in the experiments of Fig. I.2.3 (see also Table I.2.1). This can be explained as follows: in action-angle variables the Hamiltonian of the Kepler problem is $\mathcal{H}(a_1, a_2)$, where $a_2 = L$ is the angular momentum. Since the angular momentum is a quadratic invariant that is exactly conserved by symplectic integrators such as symplectic partitioned Runge–Kutta methods, the modified Hamiltonian

$$\tilde{\mathcal{H}}(a, \theta) = \mathcal{H}(a_1, a_2) + \varepsilon \mathcal{G}_h(a_1, a_2, \theta_1)$$

does not depend on the angle variable θ_2 (see Corollary IX.5.3). As in the proof of Lemma 2.1 we average out the angle θ_1 up to a certain power of ε . Since we are concerned here with one degree of freedom, the diophantine condition is trivially satisfied, and we can conclude as in Theorem 3.1.

X.4 Near-Invariant Tori on Exponentially Long Times

We refine the results for the classical perturbation series of Sect. X.2.2 to yield locally integrable behaviour, up to exponentially small deviations, over time intervals that are exponentially long in a power of the small perturbation parameter. We then combine this result with backward error analysis to show the near-preservation of invariant tori over exponentially long times in a negative power of the step size for symplectic integrators. We begin with the necessary technical estimates.

X.4.1 Estimates of Perturbation Series

We will estimate the coefficients of the perturbation series (2.5), which requires a bound for the solution of (2.6). We use the following notation: for $\rho > 0$ and with $\|\cdot\|$ the maximum norm on \mathbb{R}^d ,

$$U_\rho = \{\theta \in \mathbb{T}^d + i\mathbb{R}^d; \|\operatorname{Im} \theta\| < \rho\}$$

denotes the complex extension of the d -dimensional torus \mathbb{T}^d of width ρ . For a bounded analytic function F on U_ρ , we write

$$\|F\|_\rho = \sup_{\theta \in U_\rho} |F(\theta)|, \quad \left\| \frac{\partial F}{\partial \theta} \right\|_\rho = \sum_{j=1}^d \left\| \frac{\partial F}{\partial \theta_j} \right\|_\rho.$$

Following Arnold (1963), we prove the following bounds for the solution of the basic partial differential equation (2.2).

Lemma 4.1. *Suppose $\omega \in \mathbb{R}^d$ satisfies the diophantine condition (2.4). Let G be a bounded real-analytic function on U_ρ , and let \bar{G} denote the average of G over \mathbb{T}^d . Then, the equation*

$$\omega \cdot \frac{\partial F}{\partial \theta} + G = \bar{G}$$

has a unique real-analytic solution F on U_ρ with zero average $\bar{F} = 0$. For every positive $\delta < \min(\rho, 1)$, F is bounded on $U_{\rho-\delta}$ by

$$\|F\|_{\rho-\delta} \leq \kappa_0 \delta^{-\alpha+1} \|G\|_\rho, \quad \left\| \frac{\partial F}{\partial \theta} \right\|_{\rho-\delta} \leq \kappa_1 \delta^{-\alpha} \|G\|_\rho,$$

where $\alpha = \nu + d + 1$ and $\kappa_0 = \gamma^{-1} 8^d 2^\nu \nu!$, $\kappa_1 = \gamma^{-1} 8^d 2^{\nu+1} (\nu + 1)!$.

Rüssmann (1975, 1976) has shown that the estimates hold with the optimal exponent $\alpha = \nu + 1$ and with $\kappa_0 = 2^{d+1-\nu} \sqrt{(2\nu)!}$ and $\kappa_1 = 2^{d-\nu} \sqrt{(2\nu+2)!}$. This optimal value of α would yield slightly more favourable estimates in the following, but here we content ourselves with the simpler result given above.

Proof of Lemma 4.1. We have the Fourier series, convergent on the complex extension $\|\operatorname{Im} \theta\| < \rho$,

$$G(\theta) - \overline{G} = \sum_{k \neq 0} g_k e^{ik \cdot \theta}, \quad F(\theta) = \sum_k f_k e^{ik \cdot \theta}$$

with Fourier coefficients $f_0 = \overline{F} = 0$ and

$$f_k = -\frac{g_k}{ik \cdot \omega} \quad \text{for } k \in \mathbb{Z}^d, k \neq 0.$$

By Cauchy's estimates, $|g_k| \leq M e^{-|k|\rho}$ with $M = \|G - \overline{G}\|_\rho \leq 2\|G\|_\rho$ and $|k| = \sum |k_i|$. It follows with (2.4) that

$$\begin{aligned} \|F\|_{\rho-\delta} &\leq \sum_k |f_k| e^{|k|(\rho-\delta)} \leq \frac{M}{\gamma} \sum_k |k|^\nu e^{-|k|\delta}, \\ \left\| \frac{\partial F}{\partial \theta} \right\|_{\rho-\delta} &\leq \sum_k |f_k| \cdot |k| e^{|k|(\rho-\delta)} \leq \frac{M}{\gamma} \sum_k |k|^{\nu+1} e^{-|k|\delta}. \end{aligned}$$

It remains to bound the right-hand sums. We use the inequality $x^\nu/\nu! \leq e^x$ with $x = |k|\delta/2$ to obtain

$$\sum_k |k|^\nu e^{-|k|\delta} \leq 2^\nu \delta^{-\nu} \nu! \sum_k e^{-|k|\delta/2}.$$

The last sum is bounded by

$$\sum_k e^{-|k|\delta/2} = \left(1 + 2 \sum_{j=1}^{\infty} e^{-j\delta/2}\right)^d = \left(\frac{1 + e^{-\delta/2}}{1 - e^{-\delta/2}}\right)^d \leq (8\delta^{-1})^d.$$

Taken together, the above inequalities yield the stated bound for $\|F\|_{\rho-\delta}$. The bound for the derivative is obtained in the same way, with ν replaced by $\nu + 1$. \square

The coefficients of the perturbation series (2.5) are bounded as follows.

Lemma 4.2. *Let H_0, H_1 be real-analytic and bounded by M on the complex r -neighbourhood $B_r(b^*)$ of $b^* \in \mathbb{R}^d$ and on $B_r(b^*) \times U_\rho$, respectively. Suppose that $\omega(b^*) = (\partial H_0 / \partial a)(b^*)$ satisfies the diophantine condition (2.4). Then, the coefficients of the perturbation series (2.5) are bounded by*

$$\left\| \frac{\partial S_j}{\partial \theta}(b^*, \cdot) \right\|_{\rho/2} \leq C_0 (C_1 j^\alpha)^{j-1}$$

for all $j \geq 0$. Here $C_0 = 2r$, and $C_1 = 128(\kappa_1 M / r \rho^\alpha)^2$ with α and κ_1 of Lemma 4.1.

Proof. We recall from Sect. X.2.2 that S_j is determined by (2.6), where $K_1 = H_1$ and for $j \geq 2$,

$$\begin{aligned} K_j &= \sum_{i=2}^j \sum_{k_1+\dots+k_i=j} \frac{1}{i!} \frac{\partial^i H_0}{\partial a^i} \left(\frac{\partial S_{k_1}}{\partial \theta}, \dots, \frac{\partial S_{k_i}}{\partial \theta} \right) \\ &+ \sum_{i=1}^{j-1} \sum_{k_1+\dots+k_i=j-1} \frac{1}{i!} \frac{\partial^i H_1}{\partial a^i} \left(\frac{\partial S_{k_1}}{\partial \theta}, \dots, \frac{\partial S_{k_i}}{\partial \theta} \right). \end{aligned}$$

We fix an index, say J , set $\delta = \rho/(2J)$ and abbreviate

$$\|K_k\|_j = \|K_k(b^*, \cdot)\|_{\rho-j\delta}$$

and similarly for $\partial S_k/\partial \theta$. By (2.6) and Lemma 4.1, we have

$$\left\| \frac{\partial S_j}{\partial \theta} \right\|_j \leq \kappa_1 \delta^{-\alpha} \|K_j\|_{j-1}.$$

We use the Cauchy estimate

$$\left| \frac{1}{i!} \frac{\partial^i H_0}{\partial a^i}(v_1, \dots, v_i) \right| \leq \frac{M}{r^i} |v_1| \cdot \dots \cdot |v_i|,$$

where $|\cdot|$ denotes the sum norm on \mathbb{C}^d , and bound $\|\cdot\|_{j-1}$ by $\|\cdot\|_k$ for $k \leq j-1$. We thus obtain from the above formula for K_j

$$\begin{aligned} \|K_j\|_{j-1} &\leq \sum_{i=2}^j \sum_{k_1+\dots+k_i=j} \frac{M}{r^i} \left\| \frac{\partial S_{k_1}}{\partial \theta} \right\|_{k_1} \cdot \dots \cdot \left\| \frac{\partial S_{k_i}}{\partial \theta} \right\|_{k_i} \\ &+ \sum_{i=1}^{j-1} \sum_{k_1+\dots+k_i=j-1} \frac{M}{r^i} \left\| \frac{\partial S_{k_1}}{\partial \theta} \right\|_{k_1} \cdot \dots \cdot \left\| \frac{\partial S_{k_i}}{\partial \theta} \right\|_{k_i}. \end{aligned}$$

Combining the two bounds yields

$$\frac{1}{r} \left\| \frac{\partial S_j}{\partial \theta} \right\|_j \leq \beta_j,$$

where, with $\mu = (M/r)(\kappa_1/\delta^\alpha)$, we have $\beta_1 = \mu$ and recursively for $j \geq 2$,

$$\beta_j = \mu \sum_{i=2}^j \sum_{k_1+\dots+k_i=j} \beta_{k_1} \cdot \dots \cdot \beta_{k_i} + \mu \sum_{i=1}^{j-1} \sum_{k_1+\dots+k_i=j-1} \beta_{k_1} \cdot \dots \cdot \beta_{k_i}.$$

Multiplying this equation with ζ^j and summing over j , we see that the generating function $b(\zeta) = \sum_{j=1}^{\infty} \beta_j \zeta^j$ is given implicitly by

$$b(\zeta) - \mu\zeta = \mu \left(\frac{1}{1-b(\zeta)} - 1 - b(\zeta) \right) + \mu\zeta \left(\frac{1}{1-b(\zeta)} - 1 \right),$$

or explicitly, after solving the quadratic equation, by

$$b(\zeta) = \frac{1}{2} \frac{1}{1+\mu} - \sqrt{\frac{1}{4} \left(\frac{1}{1+\mu} \right)^2 - \frac{\mu}{1+\mu} \zeta}.$$

Hence, $b(\zeta)$ is analytic on the disc $|\zeta| < 1/(4\mu(1+\mu))$, and is there bounded by $1/(2(1+\mu))$. For $\mu \geq 1$, Cauchy's estimate yields

$$\|\partial S_j / \partial \theta\|_j \leq r \beta_j \leq 2r (8\mu^2)^{j-1}.$$

(For the uninteresting case $\mu \leq 1$ the bound is $2r \cdot 8^{j-1}$.) For $j = J$ this almost gives the stated result upon inserting the definition of μ , but with an exponent 2α instead of α . This can be reduced to α if in the above proof δ is chosen as $\delta_1 = \rho/4$ in the first step and in the other steps as $\delta_j = \rho/(4J)$. This leads to a more complicated quadratic equation where now $b(\zeta)$ is analytic for $|\zeta| \leq (C_1 J^\alpha)^{-1}$. We omit the details of this refinement of the proof. \square

For the remainder term in (2.7) we then obtain the following bound.

Lemma 4.3. *In the situation of Lemma 4.2, with $r \leq 1$ and for $C_1 N^\alpha \leq 1/(2\varepsilon)$,*

$$\|R_N(b^*, \cdot)\|_{\rho/2} \leq 4Mr \left(\frac{4C_1}{r} N^\alpha \right)^N.$$

Proof. The remainder term R_N in (2.7) is a sum of terms

$$\frac{1}{i!} \frac{\partial^i H_{k_0}}{\partial a^i} (Q_{k_1}, \dots, Q_{k_i}) \quad \text{for } k_0 + k_1 + \dots + k_i = N,$$

where

$$Q_k = \frac{\partial S_k}{\partial \theta} + \varepsilon \frac{\partial S_{k+1}}{\partial \theta} + \dots + \varepsilon^{N-k-1} \frac{\partial S_{N-1}}{\partial \theta}.$$

As long as $C_1 N^\alpha \leq 1/(2\varepsilon)$, we have, by Lemma 4.2,

$$\begin{aligned} \|Q_k(b^*, \cdot)\|_{\rho/2} &\leq \sum_{j=k}^{N-1} \varepsilon^{(j-k)} C_0 (C_1 j^\alpha)^j \\ &\leq C_0 \sum_{j=k}^{N-1} 2^{-(j-k)} \left(\frac{j}{N} \right)^{\alpha j} (C_1 N^\alpha)^k \leq 2C_0 (C_1 N^\alpha)^k. \end{aligned}$$

This implies

$$\left\| \frac{1}{i!} \frac{\partial^i H_{k_0}}{\partial a^i} (Q_{k_1}, \dots, Q_{k_i})(b^*, \cdot) \right\|_{\rho/2} \leq \frac{M}{r^i} 2C_0 (C_1 N^\alpha)^N$$

for $k_0 + k_1 + \dots + k_i = N$. (This bound is also valid when an argument different from b^* appears in the derivatives of H_0 and H_1 , as is needed for the remainder terms in the Taylor expansion.) Estimating the number of such expressions by

$$2 \sum_{i=1}^N \binom{N+i-1}{i} \leq 2 \sum_{i=0}^{2N-1} \binom{2N-1}{i} = 2^{2N}$$

yields the result. \square

X.4.2 Near-Invariant Tori of Perturbed Integrable Systems

The following result extends Lemma 2.1 to exponentially long times for sufficiently small values of the perturbation parameter.

Theorem 4.4. *Let H_0, H_1 be real-analytic on the complex r -neighbourhood $B_r(b^*)$ of $b^* \in \mathbb{R}^d$ and on $B_r(b^*) \times U_\rho$, respectively, with $r \leq 1$ and $\rho \leq 1$. Suppose that $\omega(b^*) = (\partial H_0 / \partial a)(b^*)$ satisfies the diophantine condition (2.4). There are positive constants ε_0, c_0, C such that the following holds for every positive $\beta \leq 1$ and for $\varepsilon \leq \varepsilon_0$: there exists a real-analytic symplectic change of coordinates $(a, \theta) \mapsto (b, \varphi)$ such that every solution $(b(t), \varphi(t))$ of the perturbed system in the new coordinates, starting with $\|b(0) - b^*\| \leq c_0 \varepsilon^{2\beta}$, satisfies*

$$\|b(t) - b(0)\| \leq Ct \exp(-c \varepsilon^{-\beta/\alpha}) \quad \text{for } t \leq \exp(\tfrac{1}{2} c \varepsilon^{-\beta/\alpha}).$$

Here, $\alpha = \nu + d + 1$ and $c = (16 C_1 e / r)^{-1/\alpha}$ with C_1 of Lemma 4.2. Moreover, the transformation is such that, for (a, θ) and (b, φ) related by the above coordinate transform,

$$\|a - b\| \leq C\varepsilon \quad \text{for } \|b - b^*\| \leq c_0 \varepsilon^{2\beta}, \varphi \in U_{\rho/2}.$$

The thresholds ε_0 and c_0 are such that $\varepsilon_0^{2\beta}$ is inversely proportional to γC_1^2 , and c_0 is proportional to γC_1^2 .

Remark 4.5. Theorem 4.4 is a *local* result, showing that for b_0 near b^* the tori $\{b = b_0, \varphi \in \mathbb{T}^d\}$ are nearly invariant, up to exponentially small deviations, over exponentially long times. Nekhoroshev (1977, 1979) has shown the *global* result, under a “steepness condition” which is in particular satisfied for convex Hamiltonians, that for sufficiently small ε every solution of the perturbed Hamiltonian system satisfies, for some positive constants $A, B < 1$ (proportional to the inverse of the square of the dimension),

$$\|a(t) - a(0)\| \leq \varepsilon^B \quad \text{for } t \leq \exp(\varepsilon^{-A}).$$

Remark 4.6. The constant C_1 in Lemma 4.2 and constants in similar estimates of Hamiltonian perturbation theory are very large, with the consequence that the results on the long-time behaviour derived from them are meaningful, in a rigorous sense, only for extremely small values of the perturbation parameter ε . Nevertheless, apart from their pure theoretical interest these results are of value as they describe the behaviour to be expected if one presupposes that the constants obtained from the worst-case estimations are unduly pessimistic for a given problem, as is typically the case.

Proof of Theorem 4.4. The proof combines Lemmas 4.2 and 4.3 with the proof of Lemma 2.1. An appropriate choice of the truncation indices N and m then gives the exponential estimates.

As in the proof of Lemma 2.1, we approximate $H_1(b, \theta)$ by a trigonometric polynomial of order m in θ . The error of this approximation is bounded by $\mathcal{O}(e^{-m\rho/2})$ on $B_r(b^*) \times U_{\rho/2}$, which is $\mathcal{O}(e^{-N})$ for the choice $m = 2N/\rho$ made below. By the arguments of the proof of Lemma 2.1, the estimates of Lemmas 4.2 and 4.3 (for γ replaced by $\gamma/2$, which increases C_1 to $4C_1$) are then valid in $\mathcal{O}((jm)^{-\alpha})$ and $\mathcal{O}((Nm)^{-\alpha})$ neighbourhoods of b^* : for a sufficiently small constant c^* and with $C_2 = 16C_1/r$,

$$\left\| \frac{\partial S_j}{\partial \theta}(b, \theta) \right\| \leq C_0(4C_1j^\alpha)^{j-1} \quad \text{for } \|b - b^*\| \leq c^*(jm)^{-\alpha}, \theta \in U_{\rho/2},$$

$$|R_N(b, \theta)| \leq 4Mr(C_2N^\alpha)^N \quad \text{for } \|b - b^*\| \leq c^*(Nm)^{-\alpha}, \theta \in U_{\rho/2}.$$

We now consider the symplectic change of variables $(a, \theta) \mapsto (b, \varphi)$ defined by the generating function $S(b, \theta)$. The Hamiltonian equations in the variables (b, φ) are then of the form, for $\|b - b^*\| \leq c^*(Nm)^{-\alpha}$,

$$\begin{aligned} \dot{b} &= -\frac{\partial K}{\partial \varphi}(b, \varphi) = -\varepsilon^N \frac{\partial R_N}{\partial \theta} \frac{\partial \theta}{\partial \varphi} = \mathcal{O}(\varepsilon^N (C_2N^\alpha)^N) \\ \dot{\varphi} &= \frac{\partial K}{\partial b}(b, \varphi) = \omega_{\varepsilon, N}(b) + \mathcal{O}((Nm)^\alpha \cdot \varepsilon^N (C_2N^\alpha)^N). \end{aligned} \quad (4.1)$$

Choosing $m = 2N/\rho$ and N such that $C_2N^\alpha = 1/(e\varepsilon^\beta)$ gives

$$\begin{aligned} \dot{b} &= \mathcal{O}(\exp(-c\varepsilon^{-\beta/\alpha})) \\ \dot{\varphi} &= \omega_{\varepsilon, N}(b) + \mathcal{O}(\varepsilon^{-2\beta} \exp(-c\varepsilon^{-\beta/\alpha})) \end{aligned} \quad \text{for } \|b - b^*\| \leq c_0 \varepsilon^{2\beta} \quad (4.2)$$

with $c = (C_2e)^{-\alpha}$, which yields the result. \square

X.4.3 Near-Invariant Tori of Symplectic Integrators

We return to the situation of Sect. X.3 and apply a symplectic numerical method to the integrable Hamiltonian system (3.1) with (3.2) and (3.3).

Theorem 4.7. *Consider applying a symplectic numerical integrator of order p to the real-analytic completely integrable Hamiltonian system (3.1). Suppose that $\omega(a^*)$ satisfies the diophantine condition (2.4). Then, there exist positive constants c_0, c, C and h_0 such that the following holds for all step sizes $h \leq h_0$ and for all $\mu \leq \min(p, \alpha)$ with $\alpha = \nu + d + 1$: every numerical solution starting with $\|I(p_0, q_0) - a^*\| \leq c_0 h^{2\mu}$ satisfies*

$$\|I(p_n, q_n) - I(p_0, q_0)\| \leq C h^p \quad \text{for } nh \leq \exp(c h^{-\mu/\alpha}).$$

The constants h_0, c_0, c, C depend on d, γ, ν , on bounds of the real-analytic Hamiltonian H on a complex neighbourhood of the torus $\{(p, q); I(p, q) = a^\}$, and on the numerical method.*

Proof. The proof is obtained by following the arguments of the proof of Theorem 3.1. Instead of Lemma 2.1, now Theorem 4.4 is applied to the modified Hamiltonian system (3.5) with $\varepsilon = h^p$. This gives a change of coordinates $(a, \theta) \mapsto (b, \varphi)$ $\mathcal{O}(h^p)$ -close to the identity, such that in the new variables, the solution $(\tilde{b}(t), \tilde{\varphi}(t))$ of (3.5) satisfies

$$\tilde{b}(t) = b_0 + \mathcal{O}(\exp(-ch^{-\mu/\alpha})) \quad \text{for } t \leq \exp(ch^{-\mu/\alpha}).$$

On the other hand, using the exponentially small bound of Theorem IX.7.6, together with Theorem 4.4 and the arguments of part (b) of the proof of Theorem 3.1, yields for the numerical solution in the new variables

$$b_n = \tilde{b}(t) + \mathcal{O}(\exp(-ch^{-\mu/\alpha})) \quad \text{for } t = nh \leq \exp(ch^{-\mu/\alpha}).$$

Together with $a_n - b_n = \mathcal{O}(h^p)$ this gives the result. \square

Remark 4.8. When the symplectic method is applied to a perturbed integrable system as in Theorem 4.4, then the same argument yields for $\|I(p_0, q_0) - a^*\| \leq c_0 \eta^{2\beta}$ with $\eta = \max(\varepsilon, h^p)$ and $\beta \leq 1$ the bound

$$\|I(p_n, q_n) - I(p_0, q_0)\| \leq C \eta \quad \text{for } t \leq \exp(c \eta^{-\beta/\alpha}).$$

X.5 Kolmogorov's Theorem on Invariant Tori

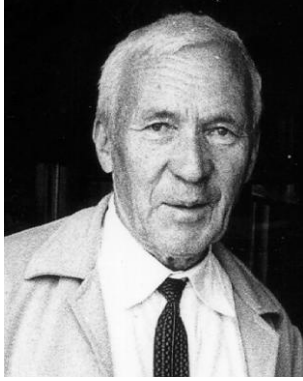
(The proof of this theorem was published in Dokl. Akad. Nauk SSSR **98** (1954), 527–530 [MR **16**, 924], but the convergence discussion does not seem convincing to the reviewer.) This very interesting theorem would imply that for an analytic canonical system which is close to an integrable one, all solutions but a set of small measure lie on invariant tori.

(J. Moser 1959)

It was a celebrated discovery by Kolmogorov (1954) that invariant tori carrying a conditionally periodic flow with diophantine frequencies persist under small perturbations of the Hamiltonian. Together with the extensions and refinements by Arnold (1963), Moser (1962) and later authors, Kolmogorov's result forms what is now known as KAM theory. Here we give a proof of Kolmogorov's theorem and use it in studying the long-time behaviour of symplectic numerical methods applied to perturbed integrable systems.

X.5.1 Kolmogorov's Theorem

In Sect. X.2.3 we have already given Kolmogorov's transformation which reduces the size of a perturbation to a Hamiltonian of the form (2.8) from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$, at least formally. The iteration of that procedure is convergent and yields the following result.

A.N. Kolmogorov²V.I. Arnold³J.K. Moser⁴

Theorem 5.1 (Kolmogorov 1954). *Consider a real-analytic Hamiltonian $H(a, \theta)$, defined for a in a neighbourhood of $0 \in \mathbb{R}^d$ and $\theta \in \mathbb{T}^d$, for which the linearization at $a^* = 0$ does not depend on the angles:*

$$H(a, \theta) = c + \omega \cdot a + \frac{1}{2} a^T M(a, \theta) a. \quad (5.1)$$

Suppose that $\omega \in \mathbb{R}^d$ satisfies the diophantine condition (2.4), viz.,

$$|k \cdot \omega| \geq \gamma |k|^{-\nu} \quad \text{for } k \in \mathbb{Z}^d, k \neq 0, \quad (5.2)$$

and that the angular average \overline{M}_0 of $M(0, \cdot)$ is an invertible $d \times d$ matrix:

$$\|\overline{M}_0 v\| \geq \mu \|v\| \quad \text{for } v \in \mathbb{R}^d, \quad (5.3)$$

with positive constants γ, ν, μ . Let $H_\varepsilon(a, \theta) = H(a, \theta) + \varepsilon G(a, \theta)$ be a real-analytic perturbation of $H(a, \theta)$. Then, there exists $\varepsilon_0 > 0$ such that for every ε with $|\varepsilon| \leq \varepsilon_0$, there is an analytic symplectic transformation $\psi_\varepsilon : (b, \varphi) \mapsto (a, \theta)$, $\mathcal{O}(\varepsilon)$ close to the identity and depending analytically on ε , which puts the perturbed Hamiltonian back to the form

$$H_\varepsilon(a, \theta) = c_\varepsilon + \omega \cdot b + \frac{1}{2} b^T M_\varepsilon(b, \varphi) b \quad \text{for } (a, \theta) = \psi_\varepsilon(b, \varphi). \quad (5.4)$$

The perturbed system therefore has the invariant torus $\{b = 0, \varphi \in \mathbb{T}^d\}$ carrying a quasi-periodic flow with the same frequencies ω as the unperturbed system. (The threshold ε_0 depends on d, ν, γ, μ and on bounds of H and G on a complex neighbourhood of $\{0\} \times \mathbb{T}^d$.)

² Andrei Nikolaevich Kolmogorov, born: 25 April 1903 in Tambov (Russia), died: 20 October 1987 in Moscow.

³ Vladimir Igorevich Arnold, born: 12 June 1937 in Odessa (USSR).

⁴ Jürgen K. Moser, born: 4 July 1928 in Königsberg, now Kaliningrad, died: 17 December 1999 in Zürich (Switzerland).

Of particular interest is the case when $H(a, \theta) = H_0(a)$ is independent of θ , so that we are considering perturbations of an integrable system. In this case, the theorem shows that all invariant tori with frequencies $\omega(a) = \partial H_0 / \partial a(a)$ satisfying (5.2) and with invertible Hessian $\partial^2 H_0 / \partial a^2(a)$ persist under small perturbations and are only slightly deformed.

Kolmogorov (1954) stated the theorem and formulated the iteration of Section X.2.3, but did not give the details of the convergence estimates. Arnold (1963) gave a first complete proof of the theorem for perturbed integrable systems, using a construction based on the “ultra-violet cutoff” (cf. Lemma 2.1) which yields a single transformation simultaneously for all frequencies satisfying the diophantine condition (2.4), in contrast to Kolmogorov's iteration which yields a different transformation for every choice of diophantine frequencies. However, Arnold's transformation is no longer analytic in the perturbation parameter ε . Moser (1962) showed that the analyticity of the Hamiltonian can be replaced by differentiability of sufficiently high order. Full proofs of Kolmogorov's theorem along his original construction were published by Thirring (1977) (for a reduced model problem) and by Benettin, Galgani, Giorgilli & Strelcyn (1984).

As in Remark 4.6, a practical difficulty with Theorem 5.1 is that the theoretically obtained threshold ε_0 is very small. The proof below requires $\varepsilon_0 \leq \delta_0^{5\alpha}$ with $\alpha = \nu + d + 1$ of Lemma 4.1, where δ_0 is inversely proportional to ν . This pessimistic estimate of the threshold can be somewhat improved by first reducing the perturbation of an integrable Hamiltonian system via a perturbation series expansion as in the proof of Theorem 4.4 and then applying Kolmogorov's theorem to the remainder of the truncated perturbation series.

The proof of Theorem 5.1 uses iteratively the following lemma, which refers to the transformation constructed in Sect. X.2.3. Similar to Sect. X.4 we use the notation

$$\|G\|_\rho = \sup\{|G(a, \theta)|; \|a\| < \rho, \|\operatorname{Im} \theta\| < \rho\}$$

for a bounded analytic function G on $W_\rho := B_\rho(0) \times U_\rho$, where again $B_\rho(0)$ is the complex ball of radius ρ around 0 and U_ρ is the complex extension of \mathbb{T}^d of width ρ . The same notation is used for vector- and matrix-valued functions, in which case the underlying norm on \mathbb{C}^d or $\mathbb{C}^{d \times d}$ is the maximum norm or its induced matrix norm, respectively.

Lemma 5.2. *In the situation of Sect. X.2.3 and under the conditions of Theorem 5.1, suppose that H and G are real-analytic and bounded on W_ρ . Then, there exists $\delta_0 > 0$ such that the following bounds hold for Kolmogorov's transformation whenever $0 < \delta \leq \delta_0$:*

$$\begin{aligned} \text{if } \|\varepsilon G\|_\rho \leq \delta^{5\alpha}, \quad \text{then } \|\varepsilon^2 \widehat{G}\|_{\rho-\delta} &\leq (\tfrac{1}{2}\delta)^{5\alpha} \\ \text{and } \|\varepsilon \nabla \chi\|_{\rho-\delta} &\leq \delta^{3\alpha}, \quad \|\widehat{M} - M\|_{\rho-\delta} \leq \delta^{2\alpha}, \end{aligned}$$

where $\alpha = \nu + d + 1$. The threshold δ_0 depends only on d, ν, γ, μ and on $\|H\|_\rho$.

Proof. We estimate the terms arising in the construction of Kolmogorov's transformation of Sect. X.2.3. For brevity we denote $\|\cdot\|_j = \|\cdot\|_{\rho-j\delta/4}$ for $j = 0, 1, 2, 3, 4$.

(a) The transformation $(b, \varphi) \mapsto (a, \theta)$ is constructed such that $(a, \theta) = y(\varepsilon)$, where $y(t)$ is the solution of $\dot{y} = J^{-1}\nabla\chi(y)$ with $y(0) = (b, \varphi)$. Suppose for the moment that

$$\|\varepsilon\nabla\chi\|_3 \leq \frac{1}{4}\delta. \quad (5.5)$$

Let $(b, \varphi) \in W_{\rho-\delta}$. Then, $\|y(t) - y(0)\| \leq \frac{1}{4}\delta$ for $0 \leq t \leq \varepsilon$, and in particular $\|(a, \theta) - (b, \varphi)\| \leq \frac{1}{4}\delta$. We define

$$\begin{aligned} \varepsilon^2 R(b, \varphi) &:= \left(a - b + \varepsilon \frac{\partial\chi}{\partial\varphi}(b, \varphi), \theta - \varphi - \varepsilon \frac{\partial\chi}{\partial b}(b, \varphi) \right) \\ &= y(\varepsilon) - y(0) - \varepsilon J^{-1}\nabla\chi(y(0)) \end{aligned}$$

and note

$$\|R(b, \varphi)\| \leq \frac{1}{2} \max_{0 \leq t \leq \varepsilon} \|\ddot{y}(t)\| \leq \frac{1}{2} \|J^{-1}\nabla^2\chi J^{-1}\nabla\chi\|_3$$

so that

$$\|R\|_4 \leq \frac{1}{2} \|\nabla^2\chi\|_3 \|\nabla\chi\|_3. \quad (5.6)$$

(b) Tracing the construction of Sect. X.2.3, we find by Taylor expansion of $H(a, \theta)$ that the new matrix is

$$\widehat{M}(b, \varphi) = M(b, \varphi) + \varepsilon L(b, \varphi)$$

with

$$L(b, \varphi) = \sum_{i=1}^d \left(\frac{\partial M}{\partial a_i} \frac{\partial\chi}{\partial\varphi_i} - \frac{\partial M}{\partial\theta_i} \frac{\partial\chi}{\partial b_i} \right) (b, \varphi) + P(b, \varphi) + Q(b, \varphi)$$

where $P(b, \varphi)$ is symmetric with

$$b^T P(b, \varphi) b = b^T \left(M(b, \varphi) - M(0, \varphi) \right) \frac{\partial\chi}{\partial\varphi}$$

and where $Q(b, \varphi)$ is given by (2.11). It follows that

$$\|\widehat{M} - M\|_4 \leq 2\varepsilon (\|\nabla M\|_4 \|\nabla\chi\|_4 + \|\nabla^2 G\|_4). \quad (5.7)$$

From the construction of \widehat{G} we also find by simple estimates of Taylor remainders

$$\|\widehat{G}\|_4 \leq \|\nabla H\|_3 \|R\|_4 + \|\nabla G\|_3 \|\nabla\chi\|_4 + \|\nabla^2 H\|_3 \|\nabla\chi\|_4^2. \quad (5.8)$$

(c) Using Lemma 4.1 in the equations (2.12)–(2.16) defining χ of (2.10), we obtain first

$$\|\chi_0\|_1 \leq \kappa_0 \delta^{-\alpha+1} \|G_0\|_0, \quad \left\| \frac{\partial\chi_0}{\partial\varphi} \right\|_1 \leq \kappa_1 \delta^{-\alpha} \|G_0\|_0$$

and by a second application of that lemma, for $i = 1, \dots, d$,

$$\|\chi_i\|_2 \leq \kappa_0 \delta^{-\alpha+1} (\|u\|_1 + \|v\|_1 + \|G_i\|_1)$$

where, by construction of u and v ,

$$\|v\|_1 \leq \|M\|_1 \left\| \frac{\partial \chi_0}{\partial \varphi} \right\|_1, \quad \|u\|_1 \leq \|M\|_1 \mu^{-1} \left(\|v\|_1 + \sum_{j=1}^d \|G_j\|_1 \right).$$

It then follows by Cauchy's estimates that

$$\|\nabla \chi\|_3 \leq C \delta^{-2\alpha} \|G\|_0, \quad \|\nabla^2 \chi\|_3 \leq C \delta^{-2\alpha-1} \|G\|_0. \quad (5.9)$$

(d) Combining the estimates (5.6)–(5.9) and using once more Cauchy's estimates to bound derivatives of H and G yields

$$\begin{aligned} \|\varepsilon^2 \widehat{G}\|_{\rho-\delta} &\leq C \delta^{-4\alpha-1} \|\varepsilon G\|_\rho^2 \\ \|\varepsilon \nabla \chi\|_{\rho-\delta} &\leq C \delta^{-2\alpha} \|\varepsilon G\|_\rho \\ \|\widehat{M} - M\|_{\rho-\delta} &\leq C \delta^{-2\alpha-3} \|\varepsilon G\|_\rho. \end{aligned}$$

All this holds under the condition (5.5). By (5.9), this condition is satisfied if $\|\varepsilon G\|_\rho \leq \delta^{5\alpha}$ and $\delta \leq \delta_0$ with a sufficiently small δ_0 . (Tracing the above constants shows that δ_0 needs to be inversely proportional to $\kappa_1^{1/\alpha}$, or inversely proportional to ν .) This yields the stated bounds. \square

Proof of Theorem 5.1. Kolmogorov's iteration yields sequences

$$\begin{aligned} G^{(0)} &= G, G^{(1)}, G^{(2)}, \dots \\ M^{(0)} &= M, M^{(1)}, M^{(2)}, \dots \\ \chi^{(0)}, \chi^{(1)}, \chi^{(2)}, \dots \end{aligned}$$

By Lemma 5.2 they satisfy, provided that $\|\varepsilon G\|_\rho = \delta^{5\alpha}$ with $\delta \leq \delta_0$,

$$\|\varepsilon^{2^j} G^{(j)}\|_{\rho^{(j)}} \leq (2^{-j} \delta)^{5\alpha} \quad (5.10)$$

$$\|M^{(j+1)} - M^{(j)}\|_{\rho^{(j)}} \leq (2^{-j} \delta)^{2\alpha} \quad (5.11)$$

$$\|\varepsilon^{2^j} \nabla \chi^{(j)}\|_{\rho^{(j)}} \leq (2^{-j} \delta)^{3\alpha} \quad (5.12)$$

where $\rho^{(j)} = \rho - (1 + \frac{1}{2} + \dots + 2^{-j})\delta > \frac{1}{2}\rho$ for all j . Note that (5.11) implies that the inverse of $M^{(j)}$ is bounded by $2\mu^{-1}$ for all j , so that the iterative use of Lemma 5.2 is justified. The time- ε^{2^j} flow of $\chi^{(j)}$ is a symplectic transformation $\sigma_\varepsilon^{(j)}$, which by (5.12) satisfies

$$\|\sigma_\varepsilon^{(j)} - \text{id}\|_{\rho/2} \leq (2^{-j} \delta)^{3\alpha}. \quad (5.13)$$

The composed transformation

$$\psi_\varepsilon^{(j)} := \sigma_\varepsilon^{(0)} \circ \sigma_\varepsilon^{(1)} \circ \dots \circ \sigma_\varepsilon^{(j)}$$

is constructed such that

$$H(\psi_\varepsilon^{(j-1)}(b, \varphi)) = c^{(j)} + \omega \cdot b + b^T M^{(j)}(b, \varphi) b + \varepsilon^{2j} G^{(j)}(b, \varphi). \quad (5.14)$$

By (5.13), the sequence $\psi_\varepsilon^{(j)}(b, \varphi)$ converges uniformly on $W_{\rho/2} \times (-\varepsilon_0, \varepsilon_0)$ to a limit $\psi_\varepsilon(b, \varphi)$. By Weierstrass' theorem, $\psi_\varepsilon(b, \varphi)$ is analytic in $(b, \varphi, \varepsilon)$ (and in any further parameters on which M and G might possibly depend analytically). Since ψ_ε depends analytically on ε and $\psi_0 = \text{id}$, it follows that ψ_ε is $\mathcal{O}(\varepsilon)$ -close to the identity on $W_{\rho/2}$. By (5.10) and (5.14), the transformed Hamiltonian $H \circ \psi_\varepsilon$ is of the desired form (5.4). \square

X.5.2 KAM Tori under Symplectic Discretization

Consider a Hamiltonian system

$$\dot{p} = -\frac{\partial \mathcal{H}}{\partial q}(p, q), \quad \dot{q} = \frac{\partial \mathcal{H}}{\partial p}(p, q), \quad (5.15)$$

for which, in suitable coordinates (a, θ) , the Hamiltonian $\mathcal{H}(p, q) = H(a, \theta) + \varepsilon G(a, \theta)$ satisfies the conditions of Theorem 5.1. Kolmogorov's theorem yields a transformation to variables (b, φ) in terms of which

$$\mathcal{H}(p, q) = \omega \cdot b + \frac{1}{2} b^T M_\varepsilon(b, \varphi) b,$$

so that the torus $\mathcal{T}_\omega = \{b = 0, \varphi \in \mathbb{T}^d\}$ is invariant and the flow on it is quasi-periodic with frequencies ω .

For a symplectic integrator of order p applied to (5.15), backward analysis gives a modified Hamiltonian $\tilde{\mathcal{H}}(p, q)$ which is an $\mathcal{O}(h^p)$ perturbation of $\mathcal{H}(p, q)$:

$$\tilde{\mathcal{H}}(p, q) = \omega \cdot b + \frac{1}{2} b^T M_\varepsilon(b, \varphi) b + h^p \tilde{G}(b, \varphi). \quad (5.16)$$

Kolmogorov's theorem can be applied once more, yielding an invariant torus $\tilde{\mathcal{T}}_\omega$ of the modified Hamiltonian $\tilde{\mathcal{H}}(p, q)$ which again carries a quasi-periodic flow with frequencies ω . Combined with the exponentially small estimates of backward analysis for the difference between numerical solutions and the flow of the modified Hamiltonian system, this gives the following result of Hairer & Lubich (1997).

Theorem 5.3. *In the above situation, for a symplectic integrator of order p used with sufficiently small step size h , there is a modified Hamiltonian $\tilde{\mathcal{H}}$ with an invariant torus $\tilde{\mathcal{T}}_\omega$ carrying a quasi-periodic flow with frequencies ω , $\mathcal{O}(h^p)$ close to the invariant torus \mathcal{T}_ω of the original Hamiltonian \mathcal{H} , such that the difference between any numerical solution (p_n, q_n) starting on the torus $\tilde{\mathcal{T}}_\omega$ and the solution*

$(\tilde{p}(t), \tilde{q}(t))$ of the modified Hamiltonian system with the same starting values remains exponentially small in $1/h$ over exponentially long times:

$$\|(p_n, q_n) - (\tilde{p}(t), \tilde{q}(t))\| \leq C e^{-\kappa/h} \quad \text{for } t = nh \leq e^{\kappa/h}.$$

The constants C and κ are independent of n, h, ε (for h, ε sufficiently small) and of the initial value $(p_0, q_0) \in \tilde{\mathcal{T}}_\omega$.

Proof. (a) For sufficiently small h , Kolmogorov's theorem applied to (5.16) yields a change of coordinates $(b, \varphi) \mapsto (c, \psi)$, $O(h^p)$ close to the identity, which transforms the modified Hamiltonian to the form

$$\tilde{\mathcal{H}}(p, q) = \omega \cdot c + \frac{1}{2} c^T M_{\varepsilon, h}(c, \psi) c,$$

with the invariant torus $\tilde{\mathcal{T}}_\omega = \{c = 0, \psi \in \mathbb{T}^d\}$. The corresponding differential equations read in these coordinates

$$\dot{c} = u(c, \psi), \quad \dot{\psi} = \omega + v(c, \psi) \quad (5.17)$$

where $u(c, \psi) = \mathcal{O}(\|c\|^2)$ and $v(c, \psi) = \mathcal{O}(\|c\|)$, and similarly for the derivatives $\partial u / \partial c = \mathcal{O}(\|c\|)$, $\partial u / \partial \psi = \mathcal{O}(\|c\|^2)$, and $\partial v / \partial c = \mathcal{O}(1)$, $\partial v / \partial \psi = \mathcal{O}(\|c\|)$. The constants in these \mathcal{O} -terms are independent of h and ε . Let $(c(t), \psi(t))$ and $(\hat{c}(t), \hat{\psi}(t))$ be two solutions of (5.17) such that $\|c(t)\| \leq \beta$, $\|\hat{c}(t)\| \leq \beta$ (β sufficiently small) for all t under consideration. Then, an argument based on Gronwall's lemma shows that their difference is bounded over a time interval $0 \leq t \leq 1/\beta$ by

$$\begin{aligned} \|c(t) - \hat{c}(t)\| &\leq C (\|c(0) - \hat{c}(0)\| + \beta \|\psi(0) - \hat{\psi}(0)\|) \\ \|\psi(t) - \hat{\psi}(t)\| &\leq C (t \|c(0) - \hat{c}(0)\| + \|\psi(0) - \hat{\psi}(0)\|), \end{aligned} \quad (5.18)$$

for some constant C that does not depend on β, h or ε .

(b) In the following we denote $y = (p, q)$ for brevity, and more specifically, y_n denotes the numerical solution starting from any y_0 on the torus $\tilde{\mathcal{T}}_\omega$, i.e., the c -coordinate of y_0 vanishes: $c_0 = 0$. We denote by $\tilde{y}(t, s, z)$ the solution of the modified Hamiltonian system with initial value $\tilde{y}(s, s, z) = z$, and more briefly $\tilde{y}(t) = \tilde{y}(t, 0, y_0)$ the solution starting from y_0 . By Theorem IX.7.6, the local error of backward error analysis at $t_j = jh$ is bounded by

$$\|y_j - \tilde{y}(t_j, t_{j-1}, y_{j-1})\| \leq \delta := \text{Const. } h e^{-3\kappa/h}$$

for some constant κ , as long as y_j remains in a compact subset of the domain of analyticity of \mathcal{H} . We further denote the c -coordinates of $y_n, \tilde{y}(t)$ and $\tilde{y}(t, t_j, y_j)$ by $c_n, \tilde{c}(t)$ and $\tilde{c}(t, t_j, y_j)$, respectively. To apply the error propagation estimate (5.18), we assume that

$$\|\tilde{c}(t, t_j, y_j)\| \leq \beta \quad \text{for } t_j \leq t \leq 1/\beta \quad (5.19)$$

and for all j satisfying $t_j = jh \leq 1/\beta$. This assumption will be justified by induction later, and the value of β will be specified in (5.21) below. By (5.18) we thus obtain the bound

$$\|\tilde{y}(t, t_j, y_j) - \tilde{y}(t, t_{j-1}, y_{j-1})\| \leq C(1 + (t - t_j))\delta \quad \text{for } t_j \leq t \leq 1/\beta.$$

Summing up from $j = 1$ to n gives for $t_n \leq t \leq 1/\beta$ (and $t > 2$)

$$\begin{aligned} \|\tilde{y}(t, t_n, y_n) - \tilde{y}(t)\| &\leq \sum_{j=1}^n C(1 + (t - t_j))\delta \leq Ch^{-1}\delta(t_n + tt_n - t_n^2/2) \\ &< Ch^{-1}\delta t^2 \leq Ch^{-1}\delta/\beta^2. \end{aligned} \quad (5.20)$$

We now set

$$\beta = (2Ch^{-1}\delta)^{1/3}, \quad (5.21)$$

so that $Ch^{-1}\delta/\beta^2 = \beta/2$, and we obtain the desired estimate from (5.20) by putting $t = t_n$.

(c) We still have to justify the assumption (5.19). This will be done by induction. For $j = 0$ nothing needs to be shown, because $\tilde{c}(t, 0, y_0) = \tilde{c}(t) \equiv 0$ as a consequence of the fact that $\tilde{y}(t)$ stays on the invariant torus $\tilde{\mathcal{T}}_\omega = \{c = 0, \psi \in \mathbb{T}^d\}$. Suppose now that (5.19) holds for $j \leq n$. It then follows from (5.20) that

$$\|\tilde{c}(t, t_n, y_n)\| < Ch^{-1}\delta/\beta^2 = \beta/2 \quad \text{for } t_n \leq t \leq 1/\beta$$

(again because of $\tilde{c}(t) \equiv 0$). Consequently we also have

$$\|c_{n+1}\| \leq \|c_{n+1} - \tilde{c}(t_{n+1}, t_n, y_n)\| + \|\tilde{c}(t_{n+1}, t_n, y_n)\| < \delta + \beta/2 \leq \beta,$$

provided that h is sufficiently small so that $\delta \leq \beta/2$. By continuity, $\tilde{c}(t, t_{n+1}, y_{n+1})$ is bounded by β on a non-empty interval $[t_{n+1}, T_{n+1}]$. The computation of part (b) shows that $\|\tilde{c}(t, t_{n+1}, y_{n+1})\| \leq \beta/2$ on this interval. Hence, T_{n+1} can be increased until $T_{n+1} \geq 1/\beta$. This proves the estimate (5.19) for $j = n + 1$. \square

X.6 Invariant Tori of Symplectic Maps

In the preceding section, backward error analysis combined with Kolmogorov's theorem has shown that a symplectic integrator applied to a Hamiltonian system with KAM tori possesses tori that are near-invariant, up to exponentially small terms, over exponentially long times in the inverse of the step size. To obtain truly invariant tori, we need a discrete KAM theorem for perturbations of integrable near-identity maps depending on a small parameter, the step size. Such a result was recently obtained by Shang (1999, 2000), who gave a discrete Arnold-type construction. Here, we use instead a discrete-time version of Kolmogorov's iteration. This establishes the existence of invariant tori of symplectic integrators applied to integrable Hamiltonian systems or to near-integrable systems with KAM tori, for a Cantor set of non-resonant step sizes.

X.6.1 A KAM Theorem for Symplectic Near-Identity Maps

We consider a discrete-time analogue of the situation in Sections X.2.3 and X.5.1 and construct the corresponding version of Kolmogorov's iteration. Consider the symplectic map $\sigma_h : (a, \theta) \mapsto (\hat{a}, \hat{\theta})$ for a near $0 \in \mathbb{R}^d$, $\theta \in \mathbb{T}^d$ defined by

$$\hat{a} = a - h \frac{\partial S}{\partial \hat{\theta}}(a, \hat{\theta}), \quad \hat{\theta} = \theta + h \frac{\partial S}{\partial a}(a, \hat{\theta}) \quad (6.1)$$

where h is a small parameter (the step size), and $S : B_r(0) \times \mathbb{T}^d \rightarrow \mathbb{R}$ is a real-analytic generating function. If $S(a, \hat{\theta})$ has the form (cf. (2.8))

$$S(a, \hat{\theta}) = c + \omega \cdot a + \frac{1}{2} a^T M(a, \hat{\theta}) a, \quad (6.2)$$

then the associated symplectic map is of the form

$$\hat{a} = a + \mathcal{O}(h\|a\|^2), \quad \hat{\theta} = \theta + h\omega + \mathcal{O}(h\|a\|).$$

Hence, the torus $\{a = 0, \theta \in \mathbb{T}^d\}$ is invariant, and on it the map σ_h reduces to rotation by $h\omega$.

Consider now an analytic perturbation of such a generating function: $S(a, \hat{\theta}) + \varepsilon R(a, \hat{\theta})$ with a small ε . We construct a near-identity symplectic change of coordinates, via an iterative procedure similar to Kolmogorov's iteration of Sect. X.2.3, such that the generating function of the perturbed symplectic map in the new variables is again of the form (6.2) with the same ω , and hence the perturbed map has an invariant torus on which it is conjugate to rotation by $h\omega$. This holds if $h\omega$ satisfies the following diophantine condition (cf. (2.4)):

$$\left| \frac{1 - e^{-ik \cdot h\omega}}{h} \right| \geq \gamma^* |k|^{-\nu^*} \quad \text{for } k \in \mathbb{Z}^d, k \neq 0, \quad (6.3)$$

for some positive constants γ^*, ν^* ; and if the angular average \overline{M}_0 of $M(0, \cdot)$ is invertible:

$$\|\overline{M}_0 v\| \geq \mu^* \|v\| \quad \text{for } v \in \mathbb{R}^d \quad (6.4)$$

for a positive constant μ^* . As in Sect. X.2.3, we construct a symplectic transformation $(a, \theta) \mapsto (b, \varphi)$ as the time- ε flow of an auxiliary Hamiltonian of the form (2.10), viz.,

$$\chi(b, \varphi) = \xi \cdot \varphi + \chi_0(\varphi) + \sum_{i=1}^d b_i \chi_i(\varphi)$$

where $\xi \in \mathbb{R}^d$ is a constant vector, and $\chi_0, \chi_1, \dots, \chi_d$ are 2π -periodic functions. We then consider the map conjugate to the perturbed map $(a, \theta) \mapsto (\hat{a}, \hat{\theta})$ generated by $S(a, \hat{\theta}) + \varepsilon R(a, \hat{\theta})$:

$$\begin{array}{ccc} (a, \theta) & \longrightarrow & (\hat{a}, \hat{\theta}) \\ \uparrow & & \downarrow \\ (b, \varphi) & & (\hat{b}, \hat{\varphi}) \end{array}$$

We construct χ in such a way that the above composed symplectic map is generated by $\tilde{S}(b, \hat{\varphi}) + \varepsilon^2 \tilde{R}(b, \hat{\varphi})$ with \tilde{S} of the form (6.2) and both \tilde{S} and \tilde{R} real-analytic and bounded independently of ε and of h with (6.3). The map $(b, \varphi) \mapsto (\hat{b}, \hat{\varphi})$ is then of the form

$$\hat{b} = b + \mathcal{O}(h\|b\|^2) + \mathcal{O}(h\varepsilon^2), \quad \hat{\varphi} = \varphi + h\omega + \mathcal{O}(h\|b\|) + \mathcal{O}(h\varepsilon^2).$$

As an elementary calculation shows, this holds if χ satisfies for all $(b, \hat{\varphi})$ with b near 0, $\hat{\varphi} \in \mathbb{T}^d$

$$\frac{\chi(b, \hat{\varphi}) - \chi(b, \hat{\varphi} - h\omega)}{h} + b^T M(b, \hat{\varphi}) \frac{\partial \chi}{\partial \varphi}(b, \hat{\varphi} - h\omega) + R(b, \hat{\varphi}) = C_h + \mathcal{O}(\|b\|^2)$$

where C_h does not depend on $(b, \hat{\varphi})$ and ε . Writing down the Taylor expansion

$$R(b, \hat{\varphi}) = R_0(\hat{\varphi}) + \sum_{i=1}^d b_i R_i(\hat{\varphi}) + \mathcal{O}(\|b\|^2)$$

and inserting the above ansatz for χ , this condition becomes fulfilled if, with $u(\hat{\varphi}) = M(0, \hat{\varphi})\xi$ and $v(\hat{\varphi}) = M(0, \hat{\varphi})(\partial\chi_0/\partial\varphi)(\hat{\varphi} - h\omega)$,

$$\frac{\chi_0(\hat{\varphi}) - \chi_0(\hat{\varphi} - h\omega)}{h} + R_0(\hat{\varphi}) = \bar{R}_0 \quad (6.5)$$

$$\frac{\chi_i(\hat{\varphi}) - \chi_i(\hat{\varphi} - h\omega)}{h} + u_i(\hat{\varphi}) + v_i(\hat{\varphi}) + R_i(\hat{\varphi}) = \bar{u}_i + \bar{v}_i + \bar{R}_i \quad (6.6)$$

$$\bar{u}_i + \bar{v}_i + \bar{R}_i = 0 \quad (i = 1, \dots, d) \quad (6.7)$$

where the bars again denote angular averages. We note

$$\frac{\chi_0(\hat{\varphi}) - \chi_0(\hat{\varphi} - h\omega)}{h} = \sum_k \frac{1 - e^{-ik \cdot h\omega}}{h} \chi_{0,k} e^{ik \cdot \hat{\varphi}},$$

where $\chi_{0,k}$ are the Fourier coefficients of χ_0 . Under the diophantine condition (6.3), Equation (6.5) is thus solved like (2.14) under condition (2.4). Equations (6.6) are of the same type. The above system is then solved in the same way as (2.12)–(2.16), yielding that the perturbed map in the new coordinates, $(b, \varphi) \mapsto (\hat{b}, \hat{\varphi})$, is generated by

$$S^{(1)}(b, \hat{\varphi}) = c^{(1)} + \omega \cdot b + \frac{1}{2} b^T M^{(1)}(b, \hat{\varphi}) b + \varepsilon^2 R^{(1)}(b, \hat{\varphi})$$

with unchanged frequencies ω and with $M^{(1)}(b, \hat{\varphi}) = M(b, \hat{\varphi}) + \mathcal{O}(\varepsilon)$. The perturbation to the form (6.2) is thus reduced from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$. By the same arguments as in the proof of Theorem 5.1 it is shown that the iteration of this procedure converges. This proves the following discrete-time version of Kolmogorov's theorem.

Theorem 6.1. *Consider a real-analytic function $S(a, \hat{\theta})$ of the form (6.2) with (6.4), defined on a neighbourhood of $\{0\} \times \mathbb{T}^d$. Let $|h| < h_0$ (h_0 so small that (6.1) is a well-defined map) and suppose that $h\omega$ satisfies (6.3).*

Let $S_\varepsilon(a, \hat{\theta}) = S(a, \theta) + \varepsilon R(a, \hat{\theta})$ be an analytic perturbation of $S(a, \theta)$, generating a symplectic map $\sigma_{h,\varepsilon} : (a, \theta) \mapsto (\hat{a}, \hat{\theta})$ via (6.1) with S_ε in place of S .

Then, there exists $\varepsilon_0 > 0$ such that for every ε with $|\varepsilon| < \varepsilon_0$, there is an analytic symplectic transformation $\psi_{h,\varepsilon} : (b, \varphi) \mapsto (a, \theta)$, $\mathcal{O}(\varepsilon)$ close to the identity uniformly in h satisfying (6.3) and analytic in ε , such that $\psi_{h,\varepsilon}^{-1} \circ \sigma_{h,\varepsilon} \circ \psi_{h,\varepsilon} : (b, \varphi) \mapsto (\hat{b}, \hat{\varphi})$ is generated, via (6.1), by a function $S_{h,\varepsilon}^*(b, \hat{\varphi})$ which is again of the form (6.2), i.e.,

$$S_{h,\varepsilon}^*(b, \hat{\varphi}) = c_{h,\varepsilon} + \omega \cdot b + \frac{1}{2} b^T M_{h,\varepsilon}(b, \hat{\varphi}) b.$$

The perturbed map $\sigma_{h,\varepsilon}$ therefore has an invariant torus on which it is conjugate to rotation by $h\omega$.

(The threshold ε_0 depends only on $d, \nu^*, \gamma^*, \mu^*$ and on bounds of S and R on a complex neighbourhood of $\{0\} \times \mathbb{T}^d$.) \square

X.6.2 Invariant Tori of Symplectic Integrators

As a direct consequence of Theorem 6.1 we obtain the following result on invariant tori of symplectic integrators applied to KAM systems.

Theorem 6.2. *Apply a symplectic integrator of order p to a perturbed integrable system with a KAM torus \mathcal{T}_ω which carries a quasi-periodic flow with diophantine frequencies ω . Then, if the step size h is sufficiently small and satisfies the strong non-resonance condition (6.3), the numerical method has an invariant torus $\mathcal{T}_{\omega,h}$ $\mathcal{O}(h^p)$ -close to \mathcal{T}_ω , on which it is conjugate to rotation by $h\omega$.*

Proof. Theorem 6.1 applies directly, with $\varepsilon = h^p$, to the above situation. Here, the generating function $S(a, \hat{\theta})$ of the time- h flow φ_h of the Hamiltonian system with the KAM torus \mathcal{T}_ω is of the form (6.2) in the variables (a, θ) obtained by Kolmogorov's theorem. The matrix $M(a, \hat{\theta})$ in (6.2) then differs from the corresponding matrix of (2.8) by $\mathcal{O}(h)$, so that (5.3) implies (6.4). Finally, the generating function of the numerical one-step map Φ_h is an $\mathcal{O}(h^p)$ -perturbation $S(a, \hat{\theta}) + h^p R(a, \hat{\theta})$. \square

X.6.3 Strongly Non-Resonant Step Sizes

Theorem 6.2 leaves us with an interesting question: if $\omega \in \mathbb{R}^d$ is a vector of frequencies that satisfies the diophantine condition (2.4), then which step sizes h satisfy the non-resonance condition (6.3)? Here we give a lemma in the spirit of results by Shang (2000). It shows that the probability of picking an $h \in (0, h_0)$ satisfying (6.3) tends to 1 as $h_0 \rightarrow 0$.

Lemma 6.3. *Suppose $\omega \in \mathbb{R}^d$ satisfies (2.4), and let $h_0 > 0$. For any choice of positive γ^* and ν^* , the set*

$$Z(h_0) = \{h \in (0, h_0) ; h \text{ does not satisfy (6.3)}\}$$

is open and dense in $(0, h_0)$. If $\gamma^* \leq \gamma$ and $\nu^* > \nu + d + r$ with $r > 1$, then the Lebesgue measure of $Z(h_0)$ is bounded by

$$\text{measure}(Z(h_0)) \leq C \frac{\gamma^*}{\gamma} h_0^{r+1}$$

where C depends only on d, ν, ν^* and $\|\omega\|$.

Proof. It is clear from the definition that $Z(h_0)$ is open and dense in $(0, h_0)$. It remains to prove the estimate of the Lebesgue measure. For every $k \in \mathbb{Z}^d$ and $|h| \leq h_0$, there exists an integer $l = l(k, h)$ such that

$$|1 - e^{-ik \cdot h\omega}| \geq \frac{2}{\pi} |k \cdot h\omega - 2\pi l| = \frac{2}{\pi} |k \cdot \omega| \cdot \left| h - \frac{2\pi l}{|k \cdot \omega|} \right|.$$

For this l we must have, by the triangle inequality,

$$2\pi|l| \leq \pi + |k| h_0 \|\omega\|,$$

so that in case $l \neq 0$

$$\frac{1}{|k|} \leq \frac{h_0 \|\omega\|}{2\pi(|l| - \frac{1}{2})}.$$

On the other hand, $l = 0$ yields

$$\left| \frac{1 - e^{-ik \cdot h\omega}}{h} \right| \geq \frac{2}{\pi} |k \cdot \omega| \geq \frac{2}{\pi} \gamma |k|^{-\nu}$$

which implies $h \notin Z(h_0)$. Hence, h can be in $Z(h_0)$ only if there exist $k \in \mathbb{Z}^d$, $k \neq 0$ and an integer $l \neq 0$ such that

$$\begin{aligned} \left| h - \frac{2\pi l}{|k \cdot \omega|} \right| &\leq \frac{\pi}{2} \frac{|h|}{|k \cdot \omega|} \frac{\gamma^*}{|k|^{\nu^*}} \leq \frac{\pi}{2} |h| \frac{|k|^\nu}{\gamma} \frac{\gamma^*}{|k|^{\nu^*}} \\ &\leq \frac{\pi}{2} \frac{\gamma^*}{\gamma} |k|^{\nu+r-\nu^*} \left(\frac{\|\omega\|}{2\pi} \frac{1}{|l| - \frac{1}{2}} \right)^r h_0^{r+1}. \end{aligned}$$

It follows that

$$\text{measure}(Z(h_0)) \leq 2 \sum_{k \neq 0} \sum_{l \neq 0} \frac{\pi}{2} \frac{\gamma^*}{\gamma} |k|^{\nu+r-\nu^*} \left(\frac{\|\omega\|}{2\pi} \frac{1}{|l| - \frac{1}{2}} \right)^r h_0^{r+1},$$

which yields the stated result. \square

X.7 Exercises

1. Let R be a $d \times 2d$ matrix of rank d . Show that there exists a symplectic $2d \times 2d$ matrix A such that $RA = (P, Q)$ with an invertible $d \times d$ matrix P .
Hint. Consider first the case $d = 2$ and then reduce the general situation to a sequence of transformations for that case.

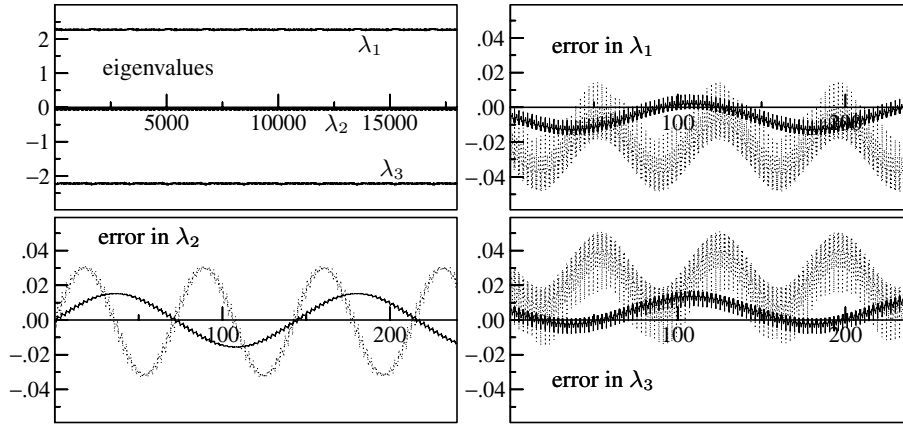


Fig. 7.1. Numerically obtained eigenvalues (left pictures) and errors in the eigenvalues (right pictures) for the step sizes $h = 0.1$ (dotted) and $h = 0.05$ (solid line)

- The transformation $(x, y) \mapsto (x, y + d(x, y))$ is symplectic if and only if the partial derivatives of d satisfy $d_x = d_x^T$, $d_y = 0$.
- In the situation of Lemma 1.1, if $(F_1, \dots, F_d, \tilde{G}_1, \dots, \tilde{G}_d)^T$ is another such symplectic transformation, then there exists a smooth function W depending only on $x = (x_1, \dots, x_d)$ such that, for $x_j = F_j(p, q)$,

$$\tilde{G}_i(p, q) - G_i(p, q) = \frac{\partial W}{\partial x_i}(x).$$

Hint. Use the previous exercise.

- Show that every discrete subgroup of \mathbb{R}^d is a grid, generated by $k \leq d$ linearly independent vectors.

Solution. See e.g. Arnold (1989), Sect. 10D.

- Show the following bound of the Lebesgue measure of non-diophantine frequencies (Arnold 1963): for any bounded domain $\Omega \subset \mathbb{R}^d$,

$$\text{measure}\{\omega \in \Omega ; \omega \text{ does not satisfy (2.4) with } \nu \geq d\} \leq C(d, \Omega)\gamma.$$

Hint. For a fixed k , decompose $\omega = \omega_0 + \alpha k/|k|$ with $\omega_0 \cdot k = 0$.

- Show that the eigenvalues λ_j of the matrix L of the Toda system are first integrals in involution.

Hint. For $P_\lambda = \det(\lambda I - L)$, show that $\{P_\lambda, P_\mu\} = 0$ for all λ, μ .

- We repeat the experiment of Fig. 1.3 with the Störmer–Verlet scheme, where we keep the initial values for the q -variables, but change the initial values for the p -variables to $p_1 = p_2 = p_3 = 0$. The numerical results, given in Fig. 7.1, are qualitatively different from those in Fig. 1.3. The errors behave more like $hc(th)$ rather than $h^2c(t)$. We do not understand this behaviour; do you?
- Show that for a non-symplectic numerical method, there is at worst quadratic error growth in time when it is applied to an integrable Hamiltonian system.

9. Consider a numerical integrator of order p (i.e., $\Phi_h(y) = \varphi_h(y) + \mathcal{O}(h^{p+1})$), and assume that

$$\Phi'_h(y)^T J \Phi'_h(y) = J + \mathcal{O}(h^{q+1})$$

with $q > p$, when the method is applied to a Hamiltonian system. Prove that under the assumptions of Theorem 3.1 the global error behaves for $t = nh$ like

$$y_n - y(t) = \mathcal{O}(th^p) + \mathcal{O}(t^2h^q),$$

and the action variables like

$$I(y_n) - I(y_0) = \mathcal{O}(h^p) + \mathcal{O}(th^q).$$

Remark. Methods satisfying the assumptions of this exercise are called *pseudo-symplectic* of order (p, q) (Aubry & Chartier 1998). Pseudo-symplectic methods behave like symplectic methods on time intervals of length $\mathcal{O}(h^{p-q})$.

10. Using the theory of B-series, in particular Theorem VI.7.4, derive the conditions for the coefficients of a Runge–Kutta method such that it is pseudo-symplectic of order $p(q)$. Prove that there exist explicit, pseudo-symplectic Runge–Kutta methods of order $(2, 4)$ with 3 stages.

Chapter XI.

Reversible Perturbation Theory and Symmetric Integrators

There is a very close similarity between the behaviour of solutions of reversible systems and that of Hamiltonian ones.

(M.B. Sevryuk 1986, p. 3)

Numerical experiments indicate that symmetric methods applied to integrable and near-integrable reversible systems share similar properties to symplectic methods applied to (near-)integrable Hamiltonian systems: linear error growth, long-time near-conservation of first integrals, existence of invariant tori. The present chapter gives a theoretical explanation of the good long-time behaviour of symmetric methods. The results and techniques are largely analogous to those of the previous chapter – the extent of the analogy may indeed be seen as the most surprising feature of this chapter.

XI.1 Integrable Reversible Systems

We consider a system of differential equations on a domain of $\mathbb{R}^m \times \mathbb{R}^n$,

$$\begin{aligned}\dot{u} &= f(u, v) \\ \dot{v} &= g(u, v),\end{aligned}\tag{1.1}$$

which is *reversible* with respect to the involution $(u, v) \mapsto (u, -v)$: for all (u, v) ,

$$\begin{aligned}f(u, -v) &= -f(u, v) \\ g(u, -v) &= g(u, v).\end{aligned}\tag{1.2}$$

From Sect. V.1 we recall that the time- t flow φ_t of a reversible system is a *reversible map*:

$$\varphi_t(u, v) = (\hat{u}, \hat{v}) \quad \text{implies} \quad \varphi_t^{-1}(u, -v) = (\hat{u}, -\hat{v}).$$

A coordinate transform $u = \mu(x, y)$, $v = \nu(x, y)$ is said to *preserve reversibility* if the relations

$$\begin{aligned}\mu(x, -y) &= \mu(x, y) \\ \nu(x, -y) &= -\nu(x, y)\end{aligned}\tag{1.3}$$

hold for all (x, y) . This implies that every reversible system (1.1) written in the new variables (x, y) is again reversible, and that every reversible map $(u, v) \mapsto (\hat{u}, \hat{v})$

expressed in the variables (x, y) again becomes a reversible map $(x, y) \mapsto (\hat{x}, \hat{y})$. Conversely, (1.3) is necessary for these properties.

For Hamiltonian systems, complete integrability is tied to the existence of a symplectic transformation to action-angle variables; see Sect. X.1. For reversible systems, we take the existence of a reversibility-preserving transformation to such variables as the definition of integrability.

Definition 1.1. The system (1.1) is called an *integrable reversible system* if, for every point $(u_0, v_0) \in \mathbb{R}^m \times \mathbb{R}^n$ in the domain of (f, g) , there exist a function $\omega = (\omega_1, \dots, \omega_n) : D \rightarrow \mathbb{R}^n$ and a diffeomorphism

$$\psi = (\mu, \nu) : D \times \mathbb{T}^n \rightarrow U \subset \mathbb{R}^m \times \mathbb{R}^n : (a, \theta) \mapsto (u, v)$$

(with D and U open sets in \mathbb{R}^m and $\mathbb{R}^m \times \mathbb{R}^n$, respectively, and $(u_0, v_0) \in U$), which preserves reversibility and transforms the system (1.1) to the form

$$\begin{aligned} \dot{a} &= 0 \\ \dot{\theta} &= \omega(a) . \end{aligned} \tag{1.4}$$

We speak of a *real-analytic integrable reversible system* if all the functions appearing in the above definition are real-analytic.

Example 1.2 (Motion in a Central Field). In Examples X.1.2 and X.1.10 we constructed action-angle variables via a series of transformations

$$\begin{pmatrix} q_1, p_2 \\ p_1, q_2 \end{pmatrix} \xrightarrow{\text{(X.1.5)}} \begin{pmatrix} r, p_\varphi \\ \varphi, p_r \end{pmatrix} \xrightarrow{\text{(X.1.6)}} \begin{pmatrix} H, L \\ y_1, y_2 \end{pmatrix} \xrightarrow{\text{(X.1.15)}} \begin{pmatrix} H, L \\ \theta_1, \theta_2 \end{pmatrix} .$$

It is easily verified that all these transformations preserve reversibility. They transform the reversible system

$$\begin{aligned} \dot{q}_1 &= p_1, & \dot{p}_2 &= -q_2 V'(r)/r \\ \dot{q}_2 &= p_2, & \dot{p}_1 &= -q_1 V'(r)/r \end{aligned} \tag{1.5}$$

(with $r = \sqrt{q_1^2 + q_2^2}$) to the form

$$\begin{aligned} \dot{H} &= 0, & \dot{L} &= 0 \\ \dot{\theta}_1 &= \frac{2\pi}{T}, & \dot{\theta}_2 &= \frac{\Phi}{T} \end{aligned} \tag{1.6}$$

with $T = T(H, L)$ and $\Phi = \Phi(H, L)$ given by (X.1.12) and (X.1.13).

As the following result shows, it is not incidental that the above transformations preserve reversibility.

Theorem 1.3. *In the situation of the Arnold–Liouville theorem, Theorem X.1.6, let the first integrals F_1, \dots, F_d of the completely integrable Hamiltonian system be such that all F_i are even functions of the second half of the arguments:*

$$F_i(u, v) = F_i(u, -v) \quad (i = 1, \dots, d). \quad (1.7)$$

Suppose that $\partial F_1/\partial u, \dots, \partial F_d/\partial u$ are linearly independent everywhere (on $\bigcup\{M_x : x \in B\}$) except possibly on a set that has no interior points. Further, assume that for every $x \in B$ there exists u such that $(u, 0) \in M_x$. Then, the transformation $\psi : (a, \theta) \mapsto (u, v)$ to action-angle variables as given by Theorem X.1.6 preserves reversibility.

Proof. The result follows by tracing the proofs of Lemma X.1.1, Theorem X.1.4 and Theorem X.1.6.

(a) For F_i satisfying (1.7) and at points where the Jacobian matrix $\partial F/\partial u$ is invertible, the construction of the local symplectic transformation $\ell = (F_1, \dots, F_d, G_1, \dots, G_d) : (u, v) \mapsto (x, y)$ shows that the generating function $S(x, v)$ becomes odd in v when the integration constant is chosen such that $S(x, 0) = 0$. By (X.1.4), this implies that ℓ preserves reversibility. A continuity argument used together with the essential uniqueness of the transformation ℓ (see Exercise X.3) does away with the exceptional points where $\partial F/\partial u$ is singular.

(b) In Theorem X.1.4, the construction of $e(x, y) = \varphi_y(\ell^{-1}(x, 0)) =: (u, v)$ is such that

$$e(x, -y) = \varphi_{-y}(\ell^{-1}(x, 0)) = (u, -v).$$

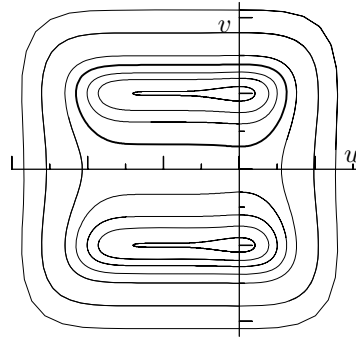
This holds because by assumption the reference point on M_x can be chosen as $\ell^{-1}(x, 0) = (u_0, 0)$ for some u_0 , and because $\varphi_{\pm y}$ is the time ± 1 flow of the Hamiltonian system with Hamiltonian $y_1 F_1 + \dots + y_d F_d$. Condition (1.7) implies that this is a reversible system, which in turn yields that e preserves reversibility as stated above.

(c) The transformation in the proof of Theorem X.1.6 is of the form $a = w(x)$, $y = W(x)\theta$ (with invertible $W(x) = w'(x)$) and hence preserves reversibility. \square

Example 1.4. We now present an example with one degree of freedom where Theorem 1.3 does not apply. In fact, all conditions are satisfied except that for some x there is no u such that $(u, 0) \in M_x$. We consider the Hamiltonian

$$H(u, v) = (v^2 - 1)^2 + \int_0^u s(s+1)^4 ds.$$

Its level sets are shown in the picture to the right. For energy values such that the level curve does not intersect the u -axis, Theorem 1.3 does not apply even though $H(u, v)$ satisfies (1.7). For these energy values the system is an integrable Hamiltonian system, but not an integrable reversible system.



Example 1.5 (Motion in a Central Field, Continued). All the assumptions of Theorem 1.3 are satisfied for $F_1 = H$, $F_2 = L = p_1 q_2 - p_2 q_1$ if we take the symplectic coordinates $u = (q_1, p_2)$ and $v = (-p_1, q_2)$.

The condition (1.7) is also satisfied with $F_1 = H$, $F_2 = L^2$ ($L \neq 0$ as always) for the choices $u = (p_1, p_2)$ and $v = (q_1, q_2)$, or $u = (q_1, q_2)$ and $v = (-p_1, -p_2)$. However, in these situations, Theorem 1.3 cannot be applied, because there does not exist u such that $(u, 0) \in M_x$.

Example 1.6 (Toda Lattice). Consider the Toda lattice of Sect. X.1.5. The eigenvalues of the matrix L are first integrals in involution. With the symplectic coordinates $(u, v) = (q, -p)$ the Hamiltonian system corresponding to (X.1.17) satisfies the reversibility conditions (1.2). However, since $v_1 + \dots + v_n$ is a first integral of this system, it is not possible to connect (u, v) with $(u, -v)$ on a level set M_x , and Theorem 1.3 cannot be applied.

Fortunately, as can be seen in Fig. 1.1, the Toda lattice contains many more symmetries. With periodic boundary conditions it is, for example, ρ -reversible (i.e., $\rho f(y) = -f(\rho y)$, $y = (p, q)^T$, see the discussion in Chap. V) with

$$\rho = \begin{pmatrix} S & 0 \\ 0 & -S \end{pmatrix} \quad S = \begin{pmatrix} & 1 \\ 1 & \end{pmatrix},$$

where S inverts the components of a vector. To bring the system to the form (1.1) with a vector field satisfying (1.2), we transform S (and hence ρ) to diagonal form and collect the variables corresponding to the eigenvalues $+1$ and -1 in u and v , respectively (see Exercise 1). This gives the (symplectic) coordinates

$$\begin{aligned} u_k &= \frac{1}{\sqrt{2}}(p_k + p_{n-k+1}), & u_{n-k+1} &= \frac{1}{\sqrt{2}}(q_k - q_{n-k+1}), \\ v_k &= \frac{1}{\sqrt{2}}(q_k + q_{n-k+1}), & v_{n-k+1} &= \frac{1}{\sqrt{2}}(p_{n-k+1} - p_k), \end{aligned} \quad (1.8)$$

for $k = 1, \dots, n/2$ (if n is even; for odd $n = 2\ell + 1$, (1.8) holds for $k = 1, \dots, \ell$ and in addition we have $u_{\ell+1} = p_{\ell+1}$ and $v_{\ell+1} = q_{\ell+1}$).

In the following we restrict our considerations to the case $n = 3$, and we show that all assumptions of Theorem 1.3 are satisfied, so that we have an integrable reversible system. For $n = 3$, the new variables are

$$\begin{aligned} u_1 &= \frac{1}{\sqrt{2}}(p_1 + p_3), & u_2 &= p_2, & u_3 &= \frac{1}{\sqrt{2}}(q_1 - q_3), \\ v_1 &= \frac{1}{\sqrt{2}}(q_1 + q_3), & v_2 &= q_2, & v_3 &= \frac{1}{\sqrt{2}}(p_3 - p_1), \end{aligned}$$

and the expressions a_k and b_k of Sect. X.1.5 become

$$\begin{aligned} a_1 &= -\frac{1}{2\sqrt{2}}(u_1 - v_3), & b_1 &= \frac{1}{2} \exp\left(\frac{1}{2}\left(\frac{1}{\sqrt{2}}(v_1 + u_3) - v_2\right)\right), \\ a_2 &= -\frac{1}{2}u_2, & b_2 &= \frac{1}{2} \exp\left(\frac{1}{2}\left(v_2 - \frac{1}{\sqrt{2}}(v_1 - u_3)\right)\right), \\ a_3 &= -\frac{1}{2\sqrt{2}}(u_1 + v_3), & b_3 &= \frac{1}{2} \exp\left(\frac{1}{\sqrt{2}}u_3\right). \end{aligned}$$

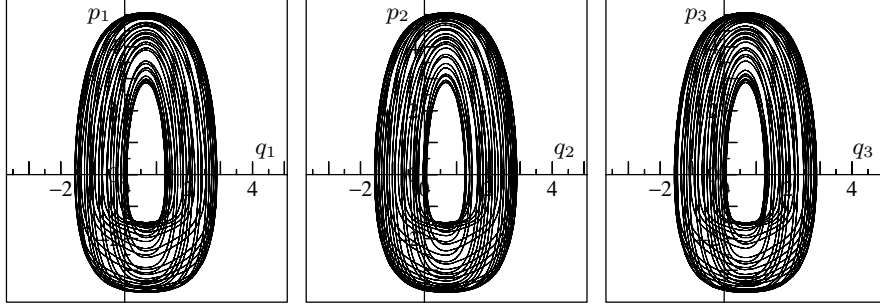


Fig. 1.1. Three projections of the solution of the Toda lattice equations ($n = 3$) with initial values as in Fig. X.1.3

One sees that $b_1^2 + b_2^2$ and $a_1 b_2^2 + a_3 b_1^2$ are even functions of v , so that all coefficients of the characteristic polynomial of the matrix L

$$\begin{aligned} \chi(\lambda) = & -\lambda^3 + (a_1 + a_2 + a_3)\lambda^2 - (a_1 a_2 + a_2 a_3 + a_3 a_1 - b_1^2 - b_2^2 - b_3^2)\lambda + \\ & (a_1 a_2 a_3 - a_1 b_2^2 - a_2 b_3^2 - a_3 b_1^2 + 2b_1 b_2 b_3). \end{aligned}$$

are even in v . This implies that also the eigenvalues of L are even functions of v , so that (1.7) is satisfied.

It remains to prove that for fixed x , i.e., for given real eigenvalues of L , the point (u_0, v_0) corresponding to $p(0), q(0)$ can be connected with an element of the form $(u, 0) \in \mathbb{R}^6$ without leaving the level set M_x . Equivalently, we have to find such a path for which the corresponding coefficients of the characteristic polynomial $\chi(\lambda)$ take given values. For given $v(t)$ this yields a system of three nonlinear equations for $u(t) \in \mathbb{R}^3$. For the eigenvalues corresponding to the initial values $p(0), q(0)$ used in Fig. X.1.3, we put $v(t) = v_0 t$ for $1 \geq t \geq 0$ and we check numerically with a path-following algorithm that such a connection is possible.

Example 1.7 (Rigid Body Equations on the Unit Sphere). We reconsider an example that has accompanied us all the way through Chapters IV, V, and VII.5: the rigid body equations (IV.1.4), here considered as differential equations on the unit sphere. We assume $I_3 < I_1, I_2$ for the inertia, which implies that any solution starting with $y_3(0) > 0$ will have $y_3(t) > 0$ for all t . We consider the equations in the neighbourhood of such a solution. We can then choose $u = y_1, v = y_2$ as coordinates on the upper half-sphere $\{y_1^2 + y_2^2 + y_3^2 = 1, y_3 > 0\}$. This gives the reversible system

$$\begin{aligned} \dot{u} &= a_1 v \sqrt{1 - u^2 - v^2} \\ \dot{v} &= a_2 u \sqrt{1 - u^2 - v^2} \end{aligned} \quad (1.9)$$

with $a_1 = (I_2 - I_3)/I_2 I_3 > 0$ and $a_2 = (I_3 - I_1)/I_3 I_1 < 0$, which has $H = u^2/I_1 + v^2/I_2 + (1 - u^2 - v^2)/I_3 = a_2 u^2 - a_1 v^2 + I_3^{-1}$ as an invariant. We introduce polar coordinates $u = r \cos \varphi, v = r \sin \varphi$ and express r as a function of H and φ :

$$r = \sqrt{\frac{I_3^{-1} - H}{a_1 \sin^2 \varphi - a_2 \cos^2 \varphi}} .$$

This leaves us with differential equations

$$\dot{H} = 0, \quad \dot{\varphi} = \gamma(H, \varphi),$$

where γ is even in φ and has no zeros. The time needed to run through an angle φ is

$$\tau(H, \varphi) = \int_0^\varphi \frac{1}{\gamma(H, \phi)} d\phi, \quad \text{and} \quad \omega(H) = \frac{2\pi}{\tau(H, 2\pi)}$$

is the frequency. With $\theta = \omega(H)\tau(H, \varphi)$ we then have

$$\dot{H} = 0, \quad \dot{\theta} = \omega(H) .$$

The transformation from (u, v) in the open unit disc (except the origin) to $(H, \theta) \in (0, I_3^{-1}) \times \mathbb{T}$ is a diffeomorphism that preserves reversibility. This shows that the rigid body equations (1.9) are an integrable reversible system.

Example 1.8 (Rigid Body Equations in \mathbb{R}^3). We now consider the rigid body equations (IV.1.4) in the ambient space \mathbb{R}^3 , rather than on the unit sphere. The system then has the invariants $H = y_1^2/I_1 + y_2^2/I_2 + y_3^2/I_3$ and $K = y_1^2 + y_2^2 + y_3^2$, and it is reversible with respect to the partition $u = (y_1, y_3)$ and $v = y_2$. In the case $I_3 < I_1, I_2$ we can again restrict our attention to $y_3 > 0$. We then write $y_3 = \sqrt{K - y_1^2 - y_2^2}$ and introduce polar coordinates $y_1 = r \cos \varphi$, $y_2 = r \sin \varphi$. As above, we express r as a function of H, K and φ (this just requires replacing I_3^{-1} with K/I_3 in the above formula for r) and we obtain differential equations

$$\dot{H} = 0, \quad \dot{K} = 0, \quad \dot{\varphi} = \gamma(H, K, \varphi)$$

with γ even in φ and without zeros. In the same way as above, this is transformed to

$$\dot{H} = 0, \quad \dot{K} = 0, \quad \dot{\theta} = \omega(H, K) .$$

The transformation $((y_1, y_3), y_2) \mapsto ((H, K), \theta)$ preserves reversibility. The rigid body equations (IV.1.4) are thus an integrable reversible system. Note that this time the dimensions differ.

XI.2 Transformations in Reversible Perturbation Theory

We consider perturbations of an integrable reversible system such that the perturbed system is still reversible. This takes the form

$$\begin{aligned}\dot{a} &= \varepsilon r(a, \theta) \\ \dot{\theta} &= \omega(a) + \varepsilon \rho(a, \theta)\end{aligned}\tag{2.1}$$

where ε is a small parameter, and r is an odd function of θ and ρ is an even function of θ :

$$\begin{aligned}r(a, -\theta) &= -r(a, \theta) \\ \rho(a, -\theta) &= \rho(a, \theta).\end{aligned}\tag{2.2}$$

Similar to Sect. X.2 for Hamiltonian perturbation theory, we study coordinate transformations that change (2.1) to reversible systems which – in various ways – look closer to an integrable system in action-angle variables than (2.1).

XI.2.1 The Basic Scheme of Reversible Perturbation Theory

We look for a transformation between neighbourhoods of $\{a_0\} \times \mathbb{T}^n$,

$$\begin{aligned}a &= b + \varepsilon s(b, \varphi) \\ \theta &= \varphi + \varepsilon \sigma(b, \varphi),\end{aligned}\tag{2.3}$$

which preserves reversibility and hence has s even in φ and σ odd in φ , such that the transformed system is of the form

$$\begin{aligned}\dot{b} &= \mathcal{O}(\varepsilon^2) \\ \dot{\varphi} &= \omega(b) + \varepsilon \mu(b) + \mathcal{O}(\varepsilon^2).\end{aligned}\tag{2.4}$$

Inserting (2.3) into (2.1) gives the system

$$\left\{ \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + \varepsilon \begin{pmatrix} \partial s / \partial b & \partial s / \partial \varphi \\ \partial \sigma / \partial b & \partial \sigma / \partial \varphi \end{pmatrix} \right\} \begin{pmatrix} \dot{b} \\ \dot{\varphi} \end{pmatrix} = \begin{pmatrix} \varepsilon r(a, \theta) \\ \omega(a) + \varepsilon \rho(a, \theta) \end{pmatrix}$$

with (a, θ) from (2.3). Inverting the matrix on the left-hand side and expanding in powers of ε , it is seen that (2.4) requires that s, σ satisfy the equations

$$\frac{\partial s}{\partial \varphi}(b, \varphi) \omega(b) = r(b, \varphi)\tag{2.5}$$

$$\frac{\partial \sigma}{\partial \varphi}(b, \varphi) \omega(b) = \rho(b, \varphi) + \omega'(b) s(b, \varphi) - \mu(b).\tag{2.6}$$

A necessary condition for the solvability of (2.5) is that the angular average of r vanishes:

$$\bar{r}(b) = 0, \quad \text{where} \quad \bar{r}(b) = \frac{1}{(2\pi)^n} \int_{\mathbb{T}^n} r(b, \varphi) d\varphi.\tag{2.7}$$

In the Hamiltonian case this condition was satisfied because r was a gradient with respect to φ . Here, in the reversible case, this is satisfied because r is an odd function of φ .

If (2.7) holds, then (2.5) can be solved by Fourier series expansion in the same way as we solved (X.2.2), provided that the frequencies $\omega_1(b), \dots, \omega_n(b)$ are non-resonant. Of course, there is again the same problem of small denominators as in the Hamiltonian case. Equations (2.6) are solved in the same way as (2.5), upon setting

$$\mu(b) = \bar{\rho}(b) + \omega'(b) \bar{s}(b). \quad (2.8)$$

Since r is odd in φ , the solution s of (2.5) becomes even in φ . It is determined uniquely only up to a constant: we are still free to choose the angular average $\bar{s}(b)$. If $\omega'(b)$ has rank n , we may actually choose $\bar{s}(b)$ such that $\mu(b) = 0$ results from (2.8). Since the right-hand side of (2.6) is even in φ , the solution σ of (2.6) becomes odd in φ if we choose $\bar{\sigma}(b) = 0$.

XI.2.2 Reversible Perturbation Series

The above construction extends to arbitrary finite order in ε . The transformation is now sought for in the form

$$a = b + \varepsilon s_1(b, \varphi) + \varepsilon^2 s_2(b, \varphi) + \dots + \varepsilon^{N-1} s_{N-1}(b, \varphi) \quad (2.9)$$

$$\theta = \varphi + \varepsilon \sigma_1(b, \varphi) + \varepsilon^2 \sigma_2(b, \varphi) + \dots + \varepsilon^{N-1} \sigma_{N-1}(b, \varphi) \quad (2.10)$$

with s_j even in φ and σ_j odd in φ to preserve reversibility. This transformation is to be chosen such that the system in the new variables is of the form

$$\begin{aligned} \dot{b} &= \varepsilon^N r_N(b, \varphi) \\ \dot{\varphi} &= \omega_{\varepsilon, N}(b) + \varepsilon^N \rho_N(b, \varphi) \end{aligned}$$

with $\omega_{\varepsilon, N}(b) = \omega(b) + \varepsilon \mu_1(b) + \dots + \varepsilon^{N-1} \mu_{N-1}(b)$, and with $r_N(b, \varphi)$ odd in φ and $\rho_N(b, \varphi)$ even in φ , and with all these functions bounded independently of ε .

Inserting the transformation into (2.1) and expanding in powers of ε , it is seen that the functions s_j and σ_j must satisfy equations of the form of (2.5), (2.6):

$$\frac{\partial s_j}{\partial \varphi}(b, \varphi) \omega(b) = p_j(b, \varphi) \quad (2.11)$$

$$\frac{\partial \sigma_j}{\partial \varphi}(b, \varphi) \omega(b) = \pi_j(b, \varphi) + \omega'(b) s_j(b, \varphi) - \mu_j(b) \quad (2.12)$$

where p_j, π_j are given by expressions that depend linearly on higher-order derivatives of r, ρ and polynomially on the functions s_i, σ_i with $i < j$ and on their first-order derivatives. Using the rules

$$\begin{pmatrix} \text{even} & \text{odd} \\ \text{odd} & \text{even} \end{pmatrix} \begin{pmatrix} \text{odd} \\ \text{even} \end{pmatrix} = \begin{pmatrix} \text{odd} \\ \text{even} \end{pmatrix}$$

and

$$\frac{\partial \text{even}}{\partial \varphi} = \text{odd}, \quad \frac{\partial \text{odd}}{\partial \varphi} = \text{even},$$

it is found that p_j is odd in φ and π_j is even in φ for all j . For non-resonant frequencies $\omega(b)$, the equations (2.11), (2.12) can therefore be solved with s_j even in φ , σ_j odd in φ . If $\omega'(b)$ is invertible, we can obtain $\mu_j(b) = 0$ for all j .

Beyond these formal calculations, there is the following reversible analogue of Lemma X.2.1 in the Hamiltonian case. This result is obtained by the same “ultra-violet cut-off” argument as the earlier result.

Lemma 2.1. *Let the right-hand side functions of (2.1) be real-analytic in a neighbourhood of $\{b^*\} \times \mathbb{T}^n$ and satisfy (2.2). Suppose that $\omega(b^*)$ satisfies the diophantine condition (X.2.4). For any fixed $N \geq 2$, there are positive constants ε_0, c, C such that the following holds for $\varepsilon \leq \varepsilon_0$: there exists a real-analytic reversibility-preserving change of coordinates $(a, \theta) \mapsto (b, \varphi)$ such that every solution $(b(t), \varphi(t))$ of the perturbed system in the new coordinates, starting with $\|b(0) - b^*\| \leq c |\log \varepsilon|^{-\nu-1}$, satisfies*

$$\begin{aligned} \|b(t) - b(0)\| &\leq C t \varepsilon^N \quad \text{for } t \leq \varepsilon^{-N+1}, \\ \|\varphi(t) - \omega_{\varepsilon, N}(b(0))t - \varphi(0)\| &\leq C (t^2 + t |\log \varepsilon|^{\nu+1}) \varepsilon^N \quad \text{for } t^2 \leq \varepsilon^{-N+1}. \end{aligned}$$

Moreover, the transformation is $\mathcal{O}(\varepsilon)$ -close to the identity: $\|(a, \theta) - (b, \varphi)\| \leq C\varepsilon$ holds for (a, θ) and (b, φ) related by the above coordinate transform, for $\|b - b^*\| \leq c |\log \varepsilon|^{-\nu-1}$ and for φ in an ε -independent complex neighbourhood of \mathbb{T}^n .

The constants ε_0, c, C depend on N, n, γ, ν and on bounds of ω, r, ρ on a complex neighbourhood of $\{b^*\} \times \mathbb{T}^n$. \square

The equations determining the coefficient functions of the perturbation series are of the form to which Lemma X.4.1 applies. Therefore, that lemma is again the tool for estimating the terms in the perturbation series, similar to Sect. X.4.1. This yields a reversible analogue of Theorem X.4.4 showing near-invariance of tori (up to exponentially small terms in a negative power of ε) over time intervals that are exponentially large in a negative power of ε , with the same exponents α, β as in Theorem X.4.4.

XI.2.3 Reversible KAM Theory

For an integrable reversible system, just as for an integrable Hamiltonian system, the phase space is foliated into invariant tori on which the flow is conditionally periodic. We fix one such torus $\{a = a^*, \theta \in \mathbb{T}^n\}$ with diophantine frequencies $\omega_1, \dots, \omega_n$. For convenience we may assume $a^* = 0 \in \mathbb{R}^m$. This torus is invariant under the flow of systems of the form $\dot{a} = \mathcal{O}(\|a\|^2)$, $\dot{\theta} = \omega + \mathcal{O}(\|a\|)$, or written more explicitly,

$$\begin{aligned} \dot{a} &= \frac{1}{2} a^T K(a, \theta) a \\ \dot{\theta} &= \omega + M(a, \theta) a. \end{aligned} \tag{2.13}$$

Here, $K = [K_1, \dots, K_m]$ where each $K_i(a, \theta)$ is a symmetric $m \times m$ matrix, and $M(a, \theta)$ is an $n \times m$ matrix. The first equation is to be interpreted as $\dot{a}_i = \frac{1}{2}a^T K_i(a, \theta)a$ for the components $i = 1, \dots, m$. Consider now a perturbation of this system:

$$\begin{aligned}\dot{a} &= \frac{1}{2}a^T K(a, \theta)a + \varepsilon r(a, \theta) \\ \dot{\theta} &= \omega + M(a, \theta)a + \varepsilon \rho(a, \theta).\end{aligned}\tag{2.14}$$

For the reversible case, i.e., for K and r odd in θ and for M and ρ even in θ , we construct a sequence of reversibility-preserving transformations in the spirit of Kolmogorov's transformation of Sect. X.2.3, which transform (2.14) back to the form (2.13) in the new variables, showing the persistence of an invariant torus with frequencies ω_i under small reversible perturbations of the system. This holds again under the diophantine condition (X.2.4) on ω and additionally under the condition that the angular average \bar{M}_0 of M at $a = 0$ has rank n . A result of this type – a reversible KAM theorem – was shown by Moser (1973), Chap. V, in a different setting. See also Sevryuk (1986) for further results in that direction.

We look for a transformation of the form

$$\begin{aligned}a &= b + \varepsilon(s(\varphi) + S(\varphi)b) \\ \theta &= \varphi + \varepsilon\sigma(\varphi)\end{aligned}\tag{2.15}$$

with an $m \times m$ matrix $S(\varphi)$. Preserving reversibility requires that s and S are even functions and σ is odd. Higher-order terms in b play no role and are therefore omitted from the beginning. We insert this into (2.14) and obtain

$$\begin{aligned}\dot{b} &= \frac{1}{2}b^T K(b, \varphi)b + \varepsilon\left\{r(0, \varphi) - \frac{\partial s}{\partial \varphi}(\varphi)\omega\right. \\ &\quad \left.+ \frac{\partial r}{\partial b}(0, \varphi)b - \frac{\partial s}{\partial \varphi}(\varphi)M(0, \varphi)b - \frac{\partial}{\partial \varphi}(S(\varphi)b)\omega + s(\varphi)^T K(0, \varphi)b\right\} \\ &\quad + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon\|b\|^2) \\ \dot{\varphi} &= \omega + M(b, \varphi)b \\ &\quad + \varepsilon\left\{\rho(0, \varphi) - \frac{\partial \sigma}{\partial \varphi}(\varphi)\omega + M(0, \varphi)s(\varphi)\right\} + \mathcal{O}(\varepsilon^2) + \mathcal{O}(\varepsilon\|b\|).\end{aligned}$$

We require that the terms in curly brackets vanish. This holds if the following equations are satisfied (the last equation is written component-wise for notational clarity):

$$\begin{aligned}\frac{\partial s}{\partial \varphi}(\varphi)\omega &= r(0, \varphi) \\ \frac{\partial \sigma}{\partial \varphi}(\varphi)\omega &= \rho(0, \varphi) + M(0, \varphi)s(\varphi) \\ \frac{\partial S_{ij}}{\partial \varphi}(\varphi)\omega &= \frac{\partial r_i}{\partial b_j}(\varphi) - \sum_k \frac{\partial s_i}{\partial \varphi_k}(\varphi)M_{kj}(0, \varphi) + \sum_k s_k(\varphi)K_{i,kj}(0, \varphi).\end{aligned}\tag{2.16}$$

Since r is odd in φ , the first equation can be solved for s even in φ , uniquely up to a constant, the angular average \bar{s} . Since the angular average of M is assumed to be of full rank n , \bar{s} can be chosen such that the angular average of the right-hand side of the equation for σ becomes zero. Since the right-hand side is even, the equation can then be solved uniquely for an odd σ . The equations for S have an odd right-hand side and can therefore be solved for an even S .

In this way, the perturbation to the form (2.13) is reduced from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$. By the same arguments as in the Hamiltonian case (see Sect. X.5), the iteration of this procedure is seen to be convergent. This finally yields a change of coordinates that preserves reversibility and transforms the perturbed system (2.14) back to the form (2.13). We summarize this in the following theorem, which is the reversible analogue of Kolmogorov's Theorem X.5.1.

Theorem 2.2. *Consider a real-analytic reversible system (2.13). Suppose that $\omega \in \mathbb{R}^n$ satisfies the diophantine condition (X.2.4), and that the angular average of $M(0, \cdot)$ is an $n \times m$ matrix of rank n . Let (2.14) be a real-analytic reversible perturbation of the system (2.13). Then, there exists $\varepsilon_0 > 0$ (which depends on the perturbation functions only through a bound of their norms on a complex neighbourhood of $\{0\} \times \mathbb{T}^n$) such that for every ε with $|\varepsilon| \leq \varepsilon_0$, there is a real-analytic transformation $\psi_\varepsilon : (b, \varphi) \mapsto (a, \theta)$, $\mathcal{O}(\varepsilon)$ close to the identity and depending analytically on ε , which preserves reversibility and puts the perturbed system back to the form (2.13) in the new variables: $\dot{b} = \mathcal{O}(\|b\|^2)$, $\dot{\varphi} = \omega + \mathcal{O}(\|b\|)$. The perturbed system therefore has the invariant torus $\{b = 0, \varphi \in \mathbb{T}^n\}$ carrying a quasi-periodic flow with the same frequencies ω as the unperturbed system. \square*

XI.2.4 Reversible Birkhoff-Type Normalization

We show that, in the situation of diophantine frequencies ω , there is a reversibility-preserving transformation that takes a reversible system of the form (2.13) to the form

$$\begin{aligned} \dot{b} &= r_k(b, \varphi) \\ \dot{\varphi} &= \omega + \zeta_k(b) + \rho_k(b, \varphi) \end{aligned} \quad \text{with} \quad r_k, \rho_k = \mathcal{O}(\|b\|^k) \quad (2.17)$$

for arbitrary $k \geq 2$, where $\zeta_k = \bar{\rho}_1 + \dots + \bar{\rho}_{k-1}$ with the bars denoting angular averages and with $\rho_1(b, \varphi) = M(b, \varphi)b$. This implies again that the invariant torus is “very sticky”: $\|b(0)\| \leq \delta$ implies $\|b(t)\| \leq 2\delta$ for $t \leq C_k \delta^{-k+1}$. As in the Hamiltonian case, a suitable choice of k would even yield time intervals exponentially long in a negative power of δ during which solutions stay within twice the initial distance δ .

The transformation to the normal form (2.17) is constructed recursively. Suppose that in some variables (a, θ) we have, for some $k \geq 2$,

$$\begin{aligned} \dot{a} &= r_{k-1}(a, \theta) \\ \dot{\theta} &= \omega + \zeta_{k-1}(a) + \rho_{k-1}(a, \theta) \end{aligned} \quad \text{with} \quad r_{k-1}, \rho_{k-1} = \mathcal{O}(\|a\|^{k-1}).$$

Note, for $k = 2$ we have $r_1 = \mathcal{O}(\|a\|^2)$ by (2.13). We search for a transformation

$$\begin{aligned} a &= b + s(b, \varphi) \\ \theta &= \varphi + \sigma(b, \varphi) \end{aligned} \quad \text{with} \quad s, \sigma = \mathcal{O}(\|b\|^{k-1}),$$

(and $s = \mathcal{O}(\|b\|^2)$ for $k = 2$) that preserves reversibility, i.e., has s even in φ and σ odd in φ , and is such that (2.17) holds. Inserting the transformation into the above differential equation shows that this is indeed achieved if s, σ solve the following system of the form (2.5), (2.6):

$$\begin{aligned} \frac{\partial s}{\partial \varphi}(b, \varphi) \omega &= r_{k-1}(b, \varphi) \\ \frac{\partial \sigma}{\partial \varphi}(b, \varphi) \omega &= \rho_{k-1}(b, \varphi) + \zeta'_{k-1}(b) s(b, \varphi) - \mu_k(b). \end{aligned}$$

Choosing $\bar{s}(b) = 0$ leads to $\mu_k = \bar{\rho}_{k-1}$ and gives (2.17) with $\zeta_k = \zeta_{k-1} + \bar{\rho}_{k-1}$.

XI.3 Linear Error Growth and Near-Preservation of First Integrals

We now study the error behaviour of reversible methods applied to integrable reversible systems. Recall from Theorem V.1.5 that symmetric methods are reversible under the compatibility condition (V.1.4). We give an analogue of Theorem X.3.1 on the error behaviour of symplectic methods applied to integrable Hamiltonian systems. We consider an integrable reversible system (1.1) (usually not given in action-angle variables) and let $(u, v) = \psi(a, \theta)$ be the reversibility-preserving transformation to action-angle variables. The inverse transformation is denoted as

$$(a, \theta) = (I(u, v), \Theta(u, v)).$$

The following is the reversible analogue of Theorem X.3.1.

Theorem 3.1. *Consider applying a reversible numerical integrator of order p to the integrable reversible system (1.1) with real-analytic right-hand side. Suppose that $\omega(a^*)$ satisfies the diophantine condition (X.2.4). Then, there exist positive constants C, c and h_0 such that the following holds for all step sizes $h \leq h_0$: every numerical solution starting with $\|I(u_0, v_0) - a^*\| \leq c |\log h|^{-\nu-1}$ satisfies*

$$\begin{aligned} \|(u_n, v_n) - (u(t), v(t))\| &\leq C t h^p \\ \|I(u_n, v_n) - I(u_0, v_0)\| &\leq C h^p \end{aligned} \quad \text{for } t = nh \leq h^{-p}. \quad (3.1)$$

The constants h_0, c, C depend on γ, ν of (X.2.4), on the dimensions, on bounds of the real-analytic functions f, g on a complex neighbourhood of the torus $\{(u, v) : I(u, v) = a^\}$, and on the numerical method.*

Proof. The proof of Theorem X.3.1 relied on Theorem IX.3.1 and Lemma X.2.1. Using their reversible analogues Theorem IX.2.3 and Lemma 2.1 with the same arguments gives the above result for the reversible case. \square

Remark 3.2. As in the analogous remark for the Hamiltonian case, the error bounds of Theorem 3.1 also hold when the reversible method is applied to a perturbed integrable system with a perturbation parameter ε bounded by a positive power of the step size: $\varepsilon \leq Kh^\alpha$ for some $\alpha > 0$.

We consider the Hamiltonian system of Example 1.4 and apply the symmetric but non-symplectic Lobatto IIIB method with step size $h = 0.01$. In the left picture of Fig. 3.1 we choose the initial value $(u_0, v_0) = (0, 1.5)$ for which the level curve of the Hamiltonian is symmetric with respect to the u -axis and the system is an integrable reversible system. The good conservation of the Hamiltonian is in agreement with Theorem 3.1. In the right picture we choose $(u_0, v_0) = (0, 0.3)$ whose level curve is the fat line in the picture of Example 1.4 which does not intersect the u -axis. Since in this situation we do not have an integrable reversible system, Theorem 3.1 cannot be applied and we cannot expect good energy conservation.

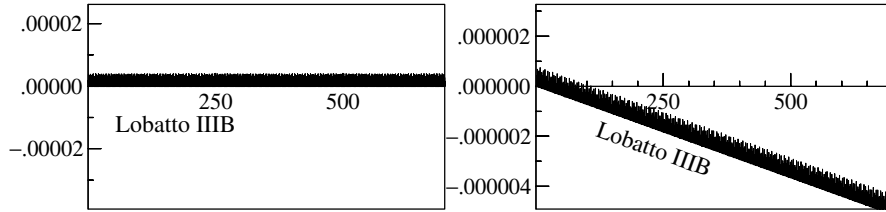


Fig. 3.1. Numerical Hamiltonian of Example 1.4 for two different initial values

For the Toda lattice example, Figures 3.2 and 3.3 illustrate the long-time conservation of the first integrals and the linear error growth, respectively, of the Lobatto IIIB method.

Theorem 3.1 together with Examples 1.7 and 1.8 also explains the good behaviour of symmetric (in fact, reversible) integrators on the rigid body equations which we observed in Chap. V (Figs. V.4.2 and V.4.6).

Variable Step Sizes: Proportional, Reversible Controllers. As a consequence of the backward error analysis of Theorem IX.6.1 the statement (3.1) can be extended straightforwardly to proportional step size controllers as discussed in Sect. VIII.3.1. Under the assumption of Theorem 3.1 with h and h_0 replaced by ε and ε_0 one has

$$\begin{aligned} \|(u_n, v_n) - (u(t_n), v(t_n))\| &\leq C t_n \varepsilon^p \\ \|I(u_n, v_n) - I(u_0, v_0)\| &\leq C \varepsilon^p \end{aligned} \quad \text{for } t_n \leq \varepsilon^{-p}. \quad (3.2)$$

The grid $\{t_n\}$ is determined by the method and satisfies $t_{n+1} = t_n + \varepsilon s(u_n, v_n, \varepsilon)$.

Variable Step Sizes: Integrating, Reversible Controllers. We apply the backward error analysis of Theorem IX.6.2. The modified equation (IX.6.14) reduces to

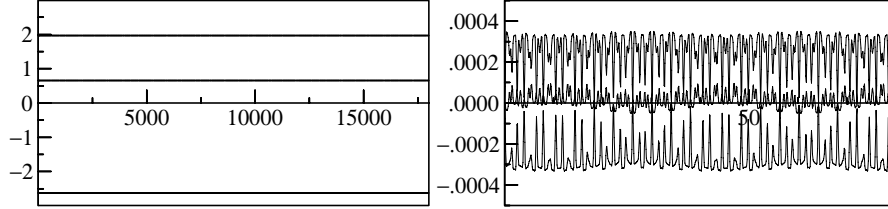


Fig. 3.2. Numerically obtained eigenvalues (left picture) and errors in the eigenvalues (right picture) of the 3-stage Lobatto IIIB scheme (step size $h = 0.1$) applied to the Toda lattice with the data of Sect. X.1.5

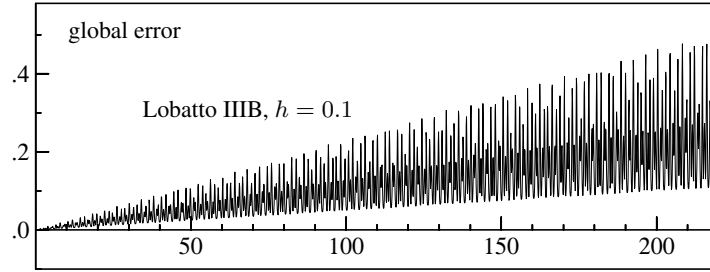


Fig. 3.3. Euclidean norm of the global error for the 3-stage Lobatto IIIB scheme (step size $h = 0.1$) applied to the Toda lattice with $n = 3$ and initial values as in Fig. 3.2

$$\dot{y} = f(y), \quad \dot{z} = z G(y) \quad (3.3)$$

for $\varepsilon = 0$. Since $G(y) = -(\sigma(y))^{-1} \nabla \sigma(y)^T f(y)$ with an analytic step size function $\sigma(y)$, the function $(y, z) \mapsto z\sigma(y)$ is a first integral of (3.3). Suppose now that $\dot{y} = f(y)$ is the integrable reversible system (1.1). This means that there exists a reversibility preserving diffeomorphism $y = \psi(a, \theta)$ transforming the system to action-angle variables. The diffeomorphism

$$\begin{pmatrix} y \\ z \end{pmatrix} = \hat{\psi}(a, A, \theta) = \begin{pmatrix} \psi(a, \theta) \\ A/\sigma(\psi(a, \theta)) \end{pmatrix}$$

is then also reversibility preserving if $\sigma(u, -v) = \sigma(u, v)$, and it transforms (3.3) to

$$\dot{a} = 0, \quad \dot{A} = 0, \quad \dot{\theta} = \omega(a).$$

If the basic method of the algorithm (IX.6.9) is reversible and if $\sigma(u, -v) = \sigma(u, v)$ holds, the modified equation (IX.6.14) is a reversible perturbation of (3.3). Consequently, Theorem 3.1 yields the statement (3.2) also for integrating step size controllers. Since $A := z\sigma(u, v)$ is an action variable, we have in addition that

$$|z_n \sigma(u_n, v_n) - z_0 \sigma(u_0, v_0)| \leq C\varepsilon^2$$

for $t_n \leq \varepsilon^{-p}$. Notice that the transformation (2.9) is $\mathcal{O}(\varepsilon^p)$ -close to the identity for the variables a and θ , but only $\mathcal{O}(\varepsilon^2)$ -close for A . This result proves that the integrating step size controller is as robust as the proportional controller. It also explains the excellent long-time behaviour observed in Figs. VIII.3.2 and VIII.3.3.

XI.4 Invariant Tori under Reversible Discretization

In this section we study the question as to how invariant tori of reversible systems are preserved under discretization of the system by reversible numerical methods. We give reversible analogues of Theorems X.5.3 and X.6.1.

XI.4.1 Near-Invariant Tori over Exponentially Long Times

We consider a reversible system (1.1) which in suitable coordinates takes the perturbed form (2.14). Under the conditions of the reversible KAM theorem, Theorem 2.2, this system has an invariant torus carrying a quasi-periodic flow with frequencies ω for sufficiently small ε . Consider now a reversible numerical integrator applied to this system. By the same arguments as in Sect. X.5.2, using the reversible KAM theorem 2.2 in place of Kolmogorov's Theorem X.5.1, we obtain the following analogue of Theorem X.5.3, which states the existence of a torus such that numerical solutions starting on this torus remain exponentially close to a quasi-periodic flow on that torus over exponentially long times in $1/h$.

Theorem 4.1. *In the above situation, for a reversible numerical method of order p used with sufficiently small step size h , there is a modified reversible system with an invariant torus $\tilde{\mathcal{T}}_\omega$ carrying a quasi-periodic flow with frequencies ω , $\mathcal{O}(h^p)$ close to the invariant torus \mathcal{T}_ω of the original reversible system, such that the difference between any numerical solution (u_n, v_n) starting on the torus $\tilde{\mathcal{T}}_\omega$ and the solution $(\tilde{u}(t), \tilde{v}(t))$ of the modified Hamiltonian system with the same starting values remains exponentially small in $1/h$ over exponentially long times:*

$$\|(u_n, v_n) - (\tilde{u}(t), \tilde{v}(t))\| \leq C e^{-\kappa/h} \quad \text{for } t = nh \leq e^{\kappa/h}.$$

The constants C and κ are independent of h, ε (for h, ε sufficiently small) and of the initial value $(u_0, v_0) \in \tilde{\mathcal{T}}_\omega$. \square

The case of initial values lying close to, but not on $\tilde{\mathcal{T}}_\omega$, can again be treated by a reversible analogue of Theorem X.4.7.

XI.4.2 A KAM Theorem for Reversible Near-Identity Maps

To obtain truly invariant tori, we need a discrete analogue of the reversible KAM theorem, which is derived in this subsection. This result can also be viewed as the reversible analogue of Theorem X.6.1. It establishes the existence of invariant tori of reversible integrators, but as in the symplectic case, only for a Cantor set of non-resonant step sizes.

A map $\Phi : (a, \theta) \mapsto (\hat{a}, \hat{\theta})$ has the invariant torus $\{a = 0, \theta \in \mathbb{T}^n\}$, and reduces on this torus to rotation by $h\omega$ (h a real parameter and $\omega \in \mathbb{R}^n$), when it is of the form (cf. (2.13))

$$\begin{aligned}\widehat{a} &= a + \frac{1}{2}ha^TK(a, \theta)a \\ \widehat{\theta} &= \theta + h\omega + hM(a, \theta)a.\end{aligned}\tag{4.1}$$

Here, $K = [K_1, \dots, K_m]$ where each $K_i(a, \theta)$ is a symmetric $m \times m$ matrix, and $M(a, \theta)$ is an $n \times m$ matrix. The expression in the first equation is again to be interpreted as $a^TK_i(a, \theta)a$ for the components $i = 1, \dots, m$.

A necessary condition for the above map Φ to be *reversible* with respect to the involution $(a, \theta) \mapsto (a, -\theta)$, cf. Definition V.1.2, is seen to be

$$\begin{aligned}K(0, -\theta) &= -K(0, \theta - h\omega) \\ M(0, -\theta) &= M(0, \theta - h\omega).\end{aligned}\tag{4.2}$$

Consider now a perturbed map

$$\begin{aligned}\widehat{a} &= a + \frac{1}{2}ha^TK(a, \theta)a + h\varepsilon r(a, \theta) \\ \widehat{\theta} &= \theta + h\omega + hM(a, \theta)a + h\varepsilon \rho(a, \theta)\end{aligned}\tag{4.3}$$

where r and ρ , which like K and M are assumed real-analytic, might depend analytically also on h and ε . Reversibility of this map implies, by direct computation, that in addition to (4.2), the following equations are satisfied up to an error $\mathcal{O}(h\varepsilon)$:

$$\begin{aligned}r(0, -\theta) &= -r(0, \theta - h\omega) \\ \frac{\partial r}{\partial a}(0, -\theta) &= -\frac{\partial r}{\partial a}(0, \theta) \\ \rho(0, -\theta) &= \rho(0, \theta - h\omega) - hM(0, \theta - h\omega)r(0, \theta - h\omega).\end{aligned}\tag{4.4}$$

Similar to Sect. XI.2.3, we construct a reversibility-preserving near-identity transformation of coordinates $(a, \theta) \mapsto (b, \varphi)$ such that the above map $\Phi_{h, \varepsilon}$ in the new variables is of the form (4.3) with the perturbation terms reduced from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$. Similar to Sect. X.6.1, this is possible if $h\omega$ satisfies the diophantine condition (X.6.3) and if the angular average \overline{M}_0 of $M(0, \cdot)$ has rank n .

We look for the transformation in the form (2.15). The functions defining this transformation must satisfy the following equations, cf. (2.16):

$$\begin{aligned}\frac{s(\varphi + h\omega) - s(\varphi)}{h} &= r(0, \varphi) \\ \frac{\sigma(\varphi + h\omega) - \sigma(\varphi)}{h} &= \rho(0, \varphi) + M(0, \varphi)s(\varphi) \\ \frac{S_{ij}(\varphi + h\omega) - S_{ij}(\varphi)}{h} &= \frac{\partial r_i}{\partial b_j}(\varphi) - \sum_k \frac{\partial s_i}{\partial \varphi_k}(\varphi)M_{kj}(0, \varphi) \\ &\quad + \sum_k s_k(\varphi)K_{i, kj}(0, \varphi).\end{aligned}\tag{4.5}$$

Under the conditions (X.6.3), (X.6.4) these equations can be solved by Fourier expansion, in the same way as the analogous equations in Sections X.6.1 and XI.2.3, and the map in the variables (b, φ) becomes of the form

$$\begin{aligned}\widehat{b} &= b + \frac{1}{2}hb^TK(b, \varphi)b + \mathcal{O}(h\varepsilon\|b\|^2) + \mathcal{O}(h\varepsilon^2) \\ \widehat{\varphi} &= \varphi + h\omega + hM(b, \varphi)b + \mathcal{O}(h\varepsilon\|b\|) + \mathcal{O}(h\varepsilon^2).\end{aligned}\quad (4.6)$$

We still need to know that the change of variables $(a, \theta) \mapsto (b, \varphi)$ preserves reversibility, i.e., that s and S are even functions of φ and σ is an odd function of φ . This is indeed a consequence of (4.2) and (4.4). (We may modify r and ρ such that (4.4) holds exactly, at the expense of introducing additional $\mathcal{O}(h^2\varepsilon^2)$ perturbations in (4.3).) Let us show this property for s . The Fourier coefficients s_k of s must satisfy

$$\frac{e^{ik \cdot h\omega} - 1}{h} s_k = r_k.$$

Since (4.4) implies $r_{-k} = -r_k e^{-ik \cdot h\omega}$ for all k , it follows that $s_{-k} = s_k$, and hence s is an even function of φ . Similarly it is shown that S is even and σ is odd.

In summary, we have found a transformation $\mathcal{O}(\varepsilon)$ close to the identity, which transforms the reversible map (4.3) to a reversible map (4.6), thus reducing the perturbation terms from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(\varepsilon^2)$. The iteration of this procedure can again be shown to be convergent. This finally yields a transformation to coordinates in terms of which the perturbed map is back in the form (2.13). In this way we obtain the following discrete analogue of Theorem 2.2 or reversible analogue of Theorem X.6.1.

Theorem 4.2. *Consider a real-analytic reversible map $\Phi_{h,\varepsilon}$ of the form (4.3), defined on a neighbourhood of $\{0\} \times \mathbb{T}^n$, with $0 \in \mathbb{R}^m$. Suppose that $h\omega$ satisfies the diophantine condition (X.6.3), and that the angular average of $M(0, \cdot)$ has rank n . Then, there exists $\varepsilon_0 > 0$ such that for every ε with $|\varepsilon| < \varepsilon_0$, there is a real-analytic transformation $\psi_{h,\varepsilon} : (b, \varphi) \mapsto (a, \theta)$, which preserves reversibility and is $\mathcal{O}(\varepsilon)$ close to the identity uniformly in h satisfying (X.6.3) and is analytic in ε , such that $\psi_{h,\varepsilon}^{-1} \circ \Phi_{h,\varepsilon} \circ \psi_{h,\varepsilon} : (b, \varphi) \mapsto (\widehat{b}, \widehat{\varphi})$ is again of the form (4.1): $\widehat{b} = b + \mathcal{O}(\|b\|^2)$, $\widehat{\varphi} = \varphi + h\omega + \mathcal{O}(\|b\|)$. The perturbed map $\Phi_{h,\varepsilon}$ therefore has an invariant torus on which it is conjugate to rotation by $h\omega$. \square*

As in the analogous situation of Sect. X.6.2, Theorem 4.2 applies directly, with $\varepsilon = h^p$, to the situation where a reversible numerical method of order p is used to discretize an integrable reversible system, or more generally, a reversible system with a KAM torus with diophantine frequencies ω . Here (4.1) corresponds to the time- h flow of the reversible system, and (4.3) represents the numerical map. This establishes the existence of invariant tori for reversible integrators, in perfect analogy to the symplectic counterpart Theorem X.6.2.

Concerning condition (X.6.3) we refer back to Sect. X.6.3, where it is shown that this condition is satisfied for a Cantor set of step sizes h if ω satisfies the diophantine condition (X.2.4).

XI.5 Exercises

1. This exercise shows that reversibility with respect to the particular involution $(u, v) \mapsto (u, -v)$ is not as special as it might seem at first glance.

- (a) If the system $\dot{y} = f(y)$ is ρ -reversible (i.e., $f(\rho y) = -\rho f(y)$), then the transformed system $\dot{z} = T^{-1}f(Tz)$ is σ -reversible with $\sigma = T^{-1}\rho T$.
- (b) Every linear involution ($\rho^2 = I$) is similar to a diagonal matrix with entries ± 1 .
2. Consider the Toda lattice equations with an arbitrary number n of degrees of freedom and with periodic boundary conditions.
- (a) Find all linear involutions ρ for which the system is ρ -reversible.
- (b) Study for which ρ the eigenvalues of the matrix L are even functions of v .
- (c) Investigate (numerically) the set of initial values for which all the assumptions of Theorem 1.3 are satisfied for some involution ρ .
- Hint.* Generalize the discussion for $n = 3$ in the Example 1.6.
3. A reversible system of the form

$$\begin{aligned}\dot{a} &= 0 \\ \dot{\theta} &= \omega(a, \theta)\end{aligned}$$

with ω an even function of $\theta \in \mathbb{T}^n$, also has a foliation of invariant tori. Consider reversible perturbations of such systems like in (2.1) and search for a reversibility-preserving transformation (2.3) that takes the perturbed system to the form

$$\begin{aligned}\dot{b} &= \mathcal{O}(\varepsilon^2) \\ \dot{\varphi} &= \omega(b, \varphi) + \varepsilon\mu(b, \varphi) + \mathcal{O}(\varepsilon^2)\end{aligned}$$

with μ even in φ . Write down the partial differential equations that the transformation must satisfy and discuss (sufficient) conditions for their solvability.

4. The torus $\{a = 0, \theta \in \mathbb{T}^n\}$ is invariant and carries a conditionally periodic flow with frequencies ω for reversible systems of the form $\dot{a} = \mathcal{O}(\|a\|)$, $\dot{\theta} = \omega + \mathcal{O}(\|a\|)$, which is more general than (2.13) in the differential equation for a . Discuss the difficulties that arise in trying to transform a reversible perturbation of such a system back to this form.
5. Apply an arbitrary (non-symmetric) Runge-Kutta method of even order $p = 2k$ to an integrable reversible system. Prove that under the assumptions of Theorem 3.1 the global error behaves for $t = nh$ like

$$y_n - y(t) = \mathcal{O}(th^p) + \mathcal{O}(t^2h^{p+1}),$$

and the action variables like

$$I(y_n) - I(y_0) = \mathcal{O}(h^p) + \mathcal{O}(th^{p+1}).$$

Chapter XII.

Dissipatively Perturbed Hamiltonian and Reversible Systems

Symplectic integrators also show a favourable long-time behaviour when they are applied to non-Hamiltonian perturbations of Hamiltonian systems. The same is true for symmetric methods applied to non-reversible perturbations of reversible systems. In this chapter we study the behaviour of numerical integrators when they are applied to dissipative perturbations of integrable systems, where only one invariant torus persists under the perturbation and becomes weakly attractive. The simplest example of such a system is Van der Pol's equation with small parameter, which has a single limit cycle in contrast to the infinitely many periodic orbits of the unperturbed harmonic oscillator.

XII.1 Numerical Experiments with Van der Pol's Equation

One of the first such methods is the method of Van-der-Pol. [...] It should, however, be noted that in the formulation given by Van-der-Pol, approximation was effected by simple intuitive reasonings.

(N.N. Bogoliubov & Y.A. Mitropolski 1961, p. 10f.)

Consider Van der Pol's equation

$$\begin{aligned}\dot{p} &= -q + \varepsilon(1 - q^2)p \\ \dot{q} &= p\end{aligned}\tag{1.1}$$

with small positive ε , which is a perturbation of the harmonic oscillator. A symplectic change to polar coordinates $p = \sqrt{2a} \cos \theta$, $q = \sqrt{2a} \sin \theta$ puts the system into the form

$$\begin{aligned}\dot{a} &= \varepsilon 2a \cos^2 \theta (1 - 2a \sin^2 \theta) \\ \dot{\theta} &= 1 + \varepsilon \cos \theta \sin \theta (1 - 2a \sin^2 \theta) .\end{aligned}$$

Since the angle θ evolves much faster than a , we may expect that the *averaged system*, which replaces the right-hand side functions by their angular averages, gives a good approximation:

$$\begin{aligned}\dot{a} &= \varepsilon a(1 - \tfrac{1}{2}a) \\ \dot{\theta} &= 1.\end{aligned}$$

Approximating by the averaged equation is the “method of Van-der-Pol” cited above, and the belief in the long-time validity of such an approximation is the *averaging principle*. The averaged differential equation for a has an unstable equilibrium at zero, and an asymptotically stable equilibrium at $a^* = 2$. The averaged system therefore has the circle $\{a^* = 2, \theta \in \mathbb{R} \bmod 2\pi\}$ as an attractive limit cycle. This suggests that the original Van der Pol equation has a nearby limit cycle, which is indeed the case.

Following the numerical experiment of Hairer & Lubich (1999), we solve the equation (1.1) with two initial values, $(p_0, q_0) = (0, 1.3)$ and $(p_0, q_0) = (0, 2.7)$, and with three numerical methods: the non-symplectic explicit and implicit Euler methods, and the symplectic Euler method. All of them have order 1. The numerical results are displayed in Fig. 1.1. For large step sizes (compared to the perturbation parameter ε), the non-symplectic methods give a completely wrong numerical solution, whereas that of the symplectic method is qualitatively correct. For smaller step sizes, the numerical solutions of the non-symplectic methods also show a limit cycle.

For the moment we explain these observations by “simple intuitive reasonings”, that is, by the averaging principle and formal backward error analysis. The rigorous treatment is developed in the course of this chapter in a more general framework of perturbed integrable systems.

For a differential equation

$$\dot{y} = f(y) + \varepsilon g(y),$$

the numerical solution y_n obtained by the explicit Euler method is the (formally) exact solution of a modified differential equation

$$\dot{\tilde{y}} = f(\tilde{y}) + \varepsilon g(\tilde{y}) - \tfrac{1}{2}h f'(\tilde{y})f(\tilde{y}) + \mathcal{O}(h^2 + \varepsilon h).$$

For the Van der Pol equation in the above coordinates, the averaged modified equation becomes

$$\dot{\tilde{a}} = h\tilde{a} + \varepsilon\tilde{a}(1 - \tfrac{1}{2}\tilde{a}) + \dots$$

which has approximately $\tilde{a} = 2 + 2h/\varepsilon$ as an equilibrium. Hence, the limit cycle of the numerical solution of the explicit Euler method has approximate radius $2\sqrt{1 + h/\varepsilon}$ (Fig. 1.1) which is far from the correct value unless $h \ll \varepsilon$.

The implicit Euler discretization is adjoint to the explicit Euler method. Therefore, its modified differential equation is as above with h replaced by $-h$. In this case, the radius of the limit cycle is approximately $2\sqrt{1 - h/\varepsilon}$ (for $h < \varepsilon$), which again agrees very well with the pictures of Fig. 1.1.

For the symplectic Euler method, the modified differential equation for Van der Pol's equation is

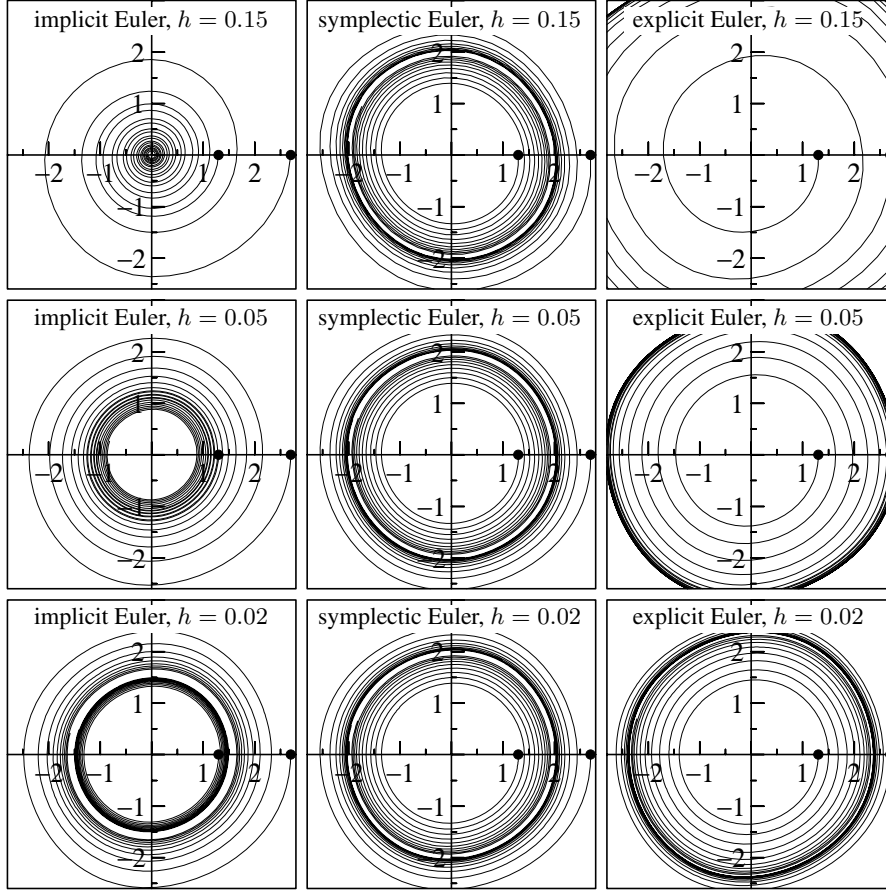


Fig. 1.1. Numerical experiments with Van der Pol's equation (1.1), $\varepsilon = 0.05$

$$\begin{aligned}\dot{\tilde{p}} &= -\tilde{q} + \varepsilon(1 - \tilde{q}^2)\tilde{p} + \frac{1}{2}h\tilde{p} + O(h^2 + \varepsilon h) \\ \dot{\tilde{q}} &= \tilde{p} - \frac{1}{2}h\tilde{q} + O(h^2 + \varepsilon h).\end{aligned}$$

Here, the modified differential equation for the unperturbed harmonic oscillator is Hamiltonian (Theorem IX.3.1), and so all ε -independent terms in the averaged modified equation vanish:

$$\int_0^{2\pi} \frac{\partial H_j}{\partial \theta}(a, \theta) d\theta = 0.$$

Therefore, the radius of the limit cycle is of size $2 + \mathcal{O}(h)$ in accordance with Fig. 1.1.

XII.2 Averaging Transformations

Le problème des oscillations non linéaires a actuellement une grande importance dans les domaines les plus divers de la technique et de la physique. Parmi les méthodes analytiques d'étude des oscillations non linéaires, la méthode asymptotique de développement en série par rapport à un paramètre petit est particulièrement efficace. Toute une série de monographies publiées en 1930–1938 par N. Krylov et N. Bogolioubov tant en russe qu'en français ont été consacrées à cette question, malheureusement ces ouvrages sont devenus aujourd'hui des raretés bibliographiques. Par ailleurs les méthodes exposées ont été largement développées depuis.

(N. Bogolioubov & I. Mitropolski 1962, préface à la traduction française)

In this section we consider rather general perturbations of integrable systems. We study transformations that eliminate the dependence on the angles in the perturbation functions, up to arbitrary powers of the small perturbation parameter. The construction and properties of these “averaging” transformations are obtained by a slight extension of the arguments in Sections X.2 and XI.2.

XII.2.1 The Basic Scheme of Averaging

As in Sections X.2.1 and XI.2.1, we consider perturbations of an integrable system written in action-angle variables:

$$\begin{aligned}\dot{a} &= \varepsilon r(a, \theta) \\ \dot{\theta} &= \omega(a) + \varepsilon \rho(a, \theta)\end{aligned}\tag{2.1}$$

where ε is a small parameter and r, ρ are real-analytic in a neighbourhood of $\{a^*\} \times \mathbb{T}^d$. Unlike the situation of the previous chapters, we do not impose conditions that make the angular average

$$\bar{r}(a) = \frac{1}{(2\pi)^d} \int_{\mathbb{T}^d} r(a, \theta) d\theta\tag{2.2}$$

vanish identically. We look for a transformation to new variables (b, φ) , of the form

$$\begin{aligned}a &= b + \varepsilon s(b, \varphi) \\ \theta &= \varphi + \varepsilon \sigma(b, \varphi),\end{aligned}\tag{2.3}$$

which eliminates the dependence on the angles in the $\mathcal{O}(\varepsilon)$ terms of (2.1):

$$\begin{aligned}\dot{b} &= \varepsilon m(b) + \mathcal{O}(\varepsilon^2) \\ \dot{\varphi} &= \omega(b) + \varepsilon \mu(b) + \mathcal{O}(\varepsilon^2).\end{aligned}\tag{2.4}$$

This is just a minor modification of the problem in Sect. XI.2.1. The equations that s and σ must satisfy, differ from (XI.2.5) and (XI.2.6) only in that the right-hand side $r(b, \varphi)$ of (XI.2.5) is replaced by $r(b, \varphi) - m(b)$, viz.,

$$\frac{\partial s}{\partial \varphi}(b, \varphi) \omega(b) = r(b, \varphi) - m(b) \quad (2.5)$$

$$\frac{\partial \sigma}{\partial \varphi}(b, \varphi) \omega(b) = \rho(b, \varphi) + \omega'(b) s(b, \varphi) - \mu(b). \quad (2.6)$$

Necessary conditions for solvability are now

$$m(b) = \bar{r}(b), \quad \mu(b) = \bar{\rho}(b), \quad (2.7)$$

where the second equation corresponds to the choice $\bar{s}(b) = 0$. In other words, the leading terms in (2.4) are the angular averages of the perturbations in (2.1).

The equations (2.5), (2.6) are solvable for $b = b^*$ if $\omega(b^*)$ satisfies the diophantine condition (X.2.4). The “ultraviolet cutoff” argument of the proof of Lemma X.2.1 then shows that (2.4) holds uniformly as long as the solution remains in the ball $\|b - b^*\| \leq c |\log \varepsilon|^{-\nu-1}$, with a sufficiently small constant c . This may hold over a very long time interval if the equation $\dot{b} = \varepsilon m(b)$ has a stable equilibrium in that ball.

XII.2.2 Perturbation Series

As in Sections X.2.2 and XI.2.2, the above construction extends to arbitrary finite order in ε . A transformation of the form (XI.2.9), which eliminates the angles in all terms up to order ε^{N-1} , is sought for:

$$\begin{aligned} \dot{b} &= \varepsilon m_1(b) + \varepsilon^2 m_2(b) + \dots + \varepsilon^{N-1} m_{N-1}(b) + \varepsilon^N r_N(b, \varphi) \\ \dot{\varphi} &= \omega(b) + \varepsilon \mu_1(b) + \varepsilon^2 \mu_2(b) + \dots + \varepsilon^{N-1} \mu_{N-1}(b) + \varepsilon^N \rho_N(b, \varphi). \end{aligned} \quad (2.8)$$

The equations determining the transformation are a slight modification of (XI.2.11) and (XI.2.12): on the right-hand side of (XI.2.11), $p_j(b, \varphi)$ is replaced by the difference $p_j(b, \varphi) - m_j(b)$, with $m_j(b) = \bar{p}_j(b)$. We then have the following variant of Lemmas X.2.1 and XI.2.1.

Lemma 2.1. *Let the right-hand side functions of (2.1) be real-analytic in a neighbourhood of $\{b^*\} \times \mathbb{T}^d$. Suppose that $\omega(b^*)$ satisfies the diophantine condition (X.2.4) with exponent ν . For any fixed $N \geq 2$, there are positive constants ε_0, c, C such that the following holds for $|\varepsilon| \leq \varepsilon_0$: there exists a real-analytic change of coordinates $(a, \theta) \mapsto (b, \varphi)$ which transforms (2.1) to (2.8) with*

$$\begin{aligned} \|m_j(b)\| &\leq C/\delta^{j-1}, & \|\mu_j(b)\| &\leq C/\delta^{j-1} \\ \|r_N(b, \varphi)\| &\leq C/\delta^{N-1}, & \|\rho_N(b, \varphi)\| &\leq C/\delta^{N-1} \end{aligned} \quad \text{for } \|b - b^*\| \leq \delta,$$

where

$$\delta = c |\log \varepsilon|^{-\nu-1}. \quad (2.9)$$

Moreover, the transformation is $\mathcal{O}(\varepsilon)$ -close to the identity: $\|(a, \theta) - (b, \varphi)\| \leq C\varepsilon$ holds for (a, θ) and (b, φ) related by the above coordinate transform, for $\|b - b^*\| \leq \delta$ and for φ in an ε -independent complex neighbourhood of \mathbb{T}^d .

The constants ε_0, c, C depend on N, d, γ, ν and on bounds of ω, r, ρ on a complex neighbourhood of $\{b^*\} \times \mathbb{T}^d$.

Proof. The proof uses again the ultraviolet cutoff argument of the proof of Lemma X.2.1. This makes all the functions $s_i, \sigma_i, m_i, \mu_i$ real-analytic in b for $\|b - b^*\| \leq 2\delta$ and of φ in an ε -independent complex neighbourhood of \mathbb{T}^d . The powers of δ in the denominators of the estimates come from the presence of terms $\partial s_j / \partial b, \partial \sigma_j / \partial b$ in $p_i(b, \varphi)$ and $\pi_i(b, \varphi)$ of (XI.2.11) and (XI.2.12) and from Cauchy's estimates applied to s_j, σ_j on $\|b - b^*\| \leq 2\delta$. \square

XII.3 Attractive Invariant Manifolds

Theorems on invariant manifolds for maps have been proved many times for many different settings. The first results were obtained by Hadamard (1901) and Perron (1929). [...] Our aim was to derive a global invariant manifold result with conditions that are easy to verify for the applications in mind. (K. Nipp & D. Stoffer 1992)

In this section we give results on the existence and properties of attractive invariant manifolds of maps, with a very explicit handling of constants. These results are due to Kirchgraber, Lasagni, Nipp & Stoffer (1991) and Nipp & Stoffer (1992). They will allow us to understand the weakly attractive closed curves that we observed in Sect. XII.1. Beyond that particular example, these results are extremely useful for studying the long-time behaviour of numerical discretizations in a great variety of applications; see Nipp & Stoffer (1995, 1996) and Lubich (2001) and references therein, and also Stuart & Humphries (1996) for a related invariant manifold theorem and its use in analyzing the dynamics of numerical integrators for non-conservative problems.

Consider a map $\Phi : X \times Y \rightarrow X \times Y$ defined on the Cartesian product of a Banach space X and a closed bounded subset Y of another Banach space. We write $\Phi(x, y) = (\hat{x}, \hat{y})$ with

$$\begin{aligned}\hat{x} &= x + f(x, y) \\ \hat{y} &= g(x, y).\end{aligned}\tag{3.1}$$

We assume that f and g are Lipschitz bounded, with Lipschitz constants L_{xx}, L_{xy} and L_{yx}, L_{yy} with respect to x, y . If these Lipschitz constants are sufficiently small, then the map Φ has an attractive invariant manifold. More precisely, there is the following result, stated without proof by Kirchgraber, Lasagni, Nipp & Stoffer (1991) and proved in a more general setting by Nipp & Stoffer (1992).

Theorem 3.1. *In the above situation, if*

$$L_{xx} + L_{yy} + 2\sqrt{L_{xy}L_{yx}} < 1, \tag{3.2}$$

then there exists a function $s : X \rightarrow Y$, which is Lipschitz bounded with the constant $\lambda = 2L_{yx}/(1 - L_{xx} - L_{yy})$, such that

$$\mathcal{M} = \{(x, s(x)) : x \in X\} \text{ is invariant under } \Phi.$$

\mathcal{M} attracts orbits of Φ with the attractivity factor $\rho = \lambda L_{xy} + L_{yy} < 1$, that is, $\|\hat{y} - s(\hat{x})\| \leq \rho \|y - s(x)\|$ holds for all $(x, y) \in X \times Y$.

Proof. (a) We search for a function $s : X \rightarrow Y$ such that for $(\hat{x}, \hat{y}) = \Phi(x, y)$, the relation $y = s(x)$ implies also $\hat{y} = s(\hat{x})$. For an arbitrary function $\sigma : X \rightarrow Y$, we first study which relation holds between \hat{x} and \hat{y} if $y = \sigma(x)$. To write \hat{y} as a function of \hat{x} , we need a bijective correspondence between x and \hat{x} via the first equation of (3.1). By the Banach fixed-point theorem, the equation

$$\hat{x} = x + f(x, \sigma(x)) \text{ has a unique solution } x = u_\sigma(\hat{x})$$

for every $\hat{x} \in X$ if $x \mapsto f(x, \sigma(x))$ is a contraction. This is the case if σ has the Lipschitz constant λ and

$$L_{xx} + L_{xy}\lambda < 1. \quad (3.3)$$

We then obtain $\hat{y} = \hat{\sigma}(\hat{x})$ from the following scheme:

$$\begin{array}{ccc} x = u_\sigma(\hat{x}) & \longleftarrow & \hat{x} \\ \downarrow \sigma & & \\ y = \sigma(x) & \longrightarrow & \hat{y} = g(x, y) \end{array}$$

That is, we set $\hat{y} = \hat{\sigma}(\hat{x}) = g(u_\sigma(\hat{x}), \sigma(u_\sigma(\hat{x})))$. By construction, $(\hat{x}, \hat{y}) = \Phi(x, y)$. Under condition (3.3), the function $u_\sigma : X \rightarrow X$ is Lipschitz bounded by $\mu = 1/(1 - L_{xx} - L_{xy}\lambda)$. Consequently, the function $\hat{\sigma} : X \rightarrow Y$ is Lipschitz bounded by $(L_{yx} + L_{yy}\lambda)\mu$. The condition that the transformed function $\hat{\sigma}$ is again Lipschitz bounded by the same λ as σ , therefore reads

$$\frac{L_{yx} + L_{yy}\lambda}{1 - L_{xx} - L_{xy}\lambda} \leq \lambda, \quad (3.4)$$

or equivalently,

$$L_{xy}\lambda^2 - (1 - L_{xx} - L_{yy})\lambda + L_{yx} \leq 0.$$

Under condition (3.2), there exists a non-empty real interval of values λ satisfying this quadratic inequality. In particular, (3.4) then holds for

$$\lambda = \frac{2L_{yx}}{1 - L_{xx} - L_{yy}}. \quad (3.5)$$

(This is close to the smallest possible value of λ if $2\sqrt{L_{xy}L_{yx}} \ll 1 - L_{xx} - L_{yy}$.) It is easily checked that (3.2) and (3.5) imply (3.3).

Under conditions (3.3) and (3.4), the transformation $H : \sigma \mapsto \hat{\sigma}$, which is called a *Hadamard graph transform*, maps the set of functions

$$S = \{\sigma : X \rightarrow Y \mid \sigma \text{ is Lipschitz bounded by } \lambda\}$$

into itself, i.e.,

$$H : S \rightarrow S : \sigma \mapsto \widehat{\sigma} .$$

S is a closed subset of $C(X, Y)$, the Banach space of continuous functions from X to the bounded closed set Y , equipped with the supremum norm $\|\sigma\|_\infty = \sup_{x \in X} \|\sigma(x)\|$. If H is a contraction, then the Banach fixed-point theorem tells us that there is a unique function $s \in S$ with $\widehat{s} = s$. By construction, this means that if $(\widehat{x}, \widehat{y}) = \Phi(x, y)$ and $y = s(x)$, then also $\widehat{y} = s(\widehat{x})$. The graph $\mathcal{M} = \{(x, s(x)) : x \in X\}$ is then an invariant manifold for the map Φ .

(b) We now show that H is already a contraction under condition (3.2). Let σ_0, σ_1 be two arbitrary functions in S , and $\widehat{x} \in X$. With $x_i = u_{\sigma_i}(\widehat{x})$,

$$\begin{aligned} \|H\sigma_1(\widehat{x}) - H\sigma_0(\widehat{x})\| &= \|g(x_1, \sigma_1(x_1)) - g(x_0, \sigma_0(x_0))\| \\ &\leq \|g(x_1, \sigma_1(x_1)) - g(x_1, \sigma_0(x_1))\| + \|g(x_1, \sigma_0(x_1)) - g(x_0, \sigma_0(x_0))\| \\ &\leq L_{yy} \|\sigma_1 - \sigma_0\|_\infty + (L_{yx} + L_{yy}\lambda) \|x_1 - x_0\| . \end{aligned}$$

By definition, $\widehat{x} = x_i + f(x_i, \sigma_i(x_i))$ for $i = 0, 1$. Subtracting these two equations yields similarly

$$\begin{aligned} \|x_1 - x_0\| &\leq \|f(x_1, \sigma_1(x_1)) - f(x_0, \sigma_0(x_0))\| \\ &\leq \|f(x_1, \sigma_1(x_1)) - f(x_1, \sigma_0(x_1))\| + \|f(x_1, \sigma_0(x_1)) - f(x_0, \sigma_0(x_0))\| \\ &\leq L_{xy} \|\sigma_1 - \sigma_0\|_\infty + (L_{xx} + L_{xy}\lambda) \|x_1 - x_0\| . \end{aligned}$$

Hence,

$$\|x_1 - x_0\| \leq \frac{L_{xy}}{1 - L_{xx} - L_{xy}\lambda} \|\sigma_1 - \sigma_0\|_\infty .$$

Combining both inequalities and recalling (3.4), we obtain

$$\|H\sigma_1 - H\sigma_0\|_\infty \leq (L_{yy} + \lambda L_{xy}) \|\sigma_1 - \sigma_0\|_\infty .$$

Since the inequality

$$L_{yy} + \lambda L_{xy} < 1 \tag{3.6}$$

is satisfied by the λ of (3.5) under condition (3.2), H is indeed a contraction.

(c) It remains to show that the invariant manifold \mathcal{M} is attractive. With $(\widehat{x}, \widehat{y}) = \Phi(x, y)$, we write

$$\begin{aligned} \widehat{y} - s(\widehat{x}) &= g(x, y) - s(x + f(x, y)) \\ &= \left(g(x, y) - g(x, s(x)) \right) + \left(s(x + f(x, s(x))) - s(x + f(x, y)) \right) . \end{aligned}$$

Here we used the identity

$$s(x + f(x, s(x))) = \widehat{s}(x + f(x, s(x))) = g(x, s(x)) ,$$

which holds because $\widehat{s} = s$ and by construction of the Hadamard transform. It follows that

$$\|\widehat{y} - s(\widehat{x})\| \leq (L_{yy} + \lambda L_{xy}) \|y - s(x)\| ,$$

which together with (3.6) yields the result. \square

Next we study the effect of a perturbation of the map on the invariant manifold.

Theorem 3.2. *Consider maps $\Phi_0, \Phi_1 : X \times Y \rightarrow X \times Y$ both of which satisfy the conditions of Theorem 3.1 with the same Lipschitz constants $L_{xx}, L_{xy}, L_{yx}, L_{yy}$. Let s_0 and s_1 be the functions defining the attractive invariant manifolds \mathcal{M}_0 and \mathcal{M}_1 , respectively. If the bound*

$$\|\Phi_1(x, y) - \Phi_0(x, y)\| \leq \delta \quad \text{for } (x, y) \in \mathcal{M}_0$$

holds in the norm $\|(x, y)\| = \lambda \|x\| + \|y\|$ on $X \times Y$, then

$$\|s_1(x) - s_0(x)\| \leq \frac{\delta}{1 - \rho} \quad \text{for } x \in X.$$

(Here λ and ρ are defined as in Theorem 3.1.)

Proof. The proof is similar to part (b) of the previous proof. Let $\hat{x} \in X$. For $i = 0, 1$, we have $s_i(\hat{x}) = g_i(x_i, s_i(x_i))$ with x_i defined by the equation $\hat{x} = x_i + f_i(x_i, s_i(x_i))$. We estimate

$$\begin{aligned} \|s_1(\hat{x}) - s_0(\hat{x})\| &\leq \|g_1(x_1, s_1(x_1)) - g_1(x_1, s_0(x_1))\| \\ &\quad + \|g_1(x_1, s_0(x_1)) - g_1(x_0, s_0(x_0))\| \\ &\quad + \|g_1(x_0, s_0(x_0)) - g_0(x_0, s_0(x_0))\| \\ &\leq L_{yy}\|s_1 - s_0\|_\infty + (L_{yx} + L_{yy}\lambda)\|x_1 - x_0\| \\ &\quad + \|g_1(x_0, s_0(x_0)) - g_0(x_0, s_0(x_0))\| \end{aligned}$$

and in the same way

$$\begin{aligned} \|x_1 - x_0\| &\leq \|f_1(x_1, s_1(x_1)) - f_1(x_1, s_0(x_1))\| \\ &\quad + \|f_1(x_1, s_0(x_1)) - f_1(x_0, s_0(x_0))\| \\ &\quad + \|f_1(x_0, s_0(x_0)) - f_0(x_0, s_0(x_0))\| \\ &\leq L_{xy}\|s_1 - s_0\|_\infty + (L_{xx} + L_{xy}\lambda)\|x_1 - x_0\| \\ &\quad + \|f_1(x_0, s_0(x_0)) - f_0(x_0, s_0(x_0))\|. \end{aligned}$$

Inserting the second bound into the first one and using (3.4) and the assumed bound on $\Phi_1 - \Phi_0$ gives

$$\|s_1 - s_0\|_\infty \leq (L_{yy} + \lambda L_{xy})\|s_1 - s_0\|_\infty + \delta,$$

which implies the result. \square

XII.4 Weakly Attractive Invariant Tori of Perturbed Integrable Systems

We assume that the perturbation is dissipative such that one torus persists under the perturbation and gets attractive.

Our analysis is done by the method of averaging. The problem of this section is classical, see e.g. Bogoliubov & Mitropolski (1961), Kirchgraber & Stiefel (1978). (D. Stoffer 1998)

In the example of the Van der Pol equation, we have seen that only one of the periodic orbits of the harmonic oscillator persists under the small nonlinear perturbation and becomes an attractive limit cycle. More generally, we consider perturbations of integrable systems

$$\begin{aligned}\dot{a} &= \varepsilon r(a, \theta) \\ \dot{\theta} &= \omega(a) + \varepsilon \rho(a, \theta)\end{aligned}\tag{4.1}$$

where (locally) just one invariant torus survives the perturbation and attracts nearby solutions. Using the results of the two previous sections, it will be shown that this situation occurs if, at some point a^* where the frequencies $\omega_i(a^*)$ are diophantine, the angular average $\bar{r}(a^*)$ is small and its Jacobian matrix

$$A = \bar{r}'(a^*)$$

has all eigenvalues with negative real part.

The following theorem is a slight modification of a result of Stoffer (1998). Early versions of it are much older; see the citations above. The origins of the problem can be traced back to the work of Van der Pol (1927) and Krylov & Bogoliubov (1934).

Here we assume the following: $\omega(a^*)$ satisfies the diophantine condition (X.2.4) with exponent ν . The perturbation functions $r(a, \theta)$ and $\rho(a, \theta)$ are real-analytic on a fixed complex neighbourhood of $\{a^*\} \times \mathbb{T}^d$ and bounded independently of ε (though they may depend on ε). In some norm $\|\cdot\|$ on \mathbb{R}^d and its induced matrix norm, the bounds

$$\|\bar{r}(a^*)\| \leq C |\log \varepsilon|^{-2(\nu+1)}\tag{4.2}$$

$$\|e^{tA}\| \leq e^{-t\alpha} \quad \text{for } t > 0\tag{4.3}$$

hold with some constants C and $\alpha > 0$.

Theorem 4.1. *Under the above conditions, for sufficiently small $\varepsilon > 0$, the system (4.1) has an invariant torus \mathcal{T}_ε which attracts an $\mathcal{O}(|\log \varepsilon|^{-\nu-1})$ -neighbourhood of $\{a^*\} \times \mathbb{T}^d$ with an exponential rate proportional to ε .*

Proof. The proof combines Lemma 2.1 and Theorem 3.1. For convenience we assume $a^* = 0$ in the following. Lemma 2.1 (with $N = 3$) gives us a change of coordinates $(a, \theta) \mapsto (b, \varphi)$, $\mathcal{O}(\varepsilon)$ -close to the identity, such that for $\|b\| \leq \delta$ with $\delta = c |\log \varepsilon|^{-\nu-1}$ of (2.9),

$$\begin{aligned}\dot{b} &= \varepsilon m_1(b) + \varepsilon^2 m_2(b) + \mathcal{O}(\varepsilon^3/\delta^2) \\ \dot{\varphi} &= \omega(b) + \varepsilon \mu_1(b) + \varepsilon^2 \mu_2(b) + \mathcal{O}(\varepsilon^3/\delta^2).\end{aligned}\quad (4.4)$$

Since $m_1(b) = \bar{r}(b) = Ab + \mathcal{O}(\delta^2)$ by (4.2), this system is of the form

$$\begin{aligned}\dot{b} &= \varepsilon Ab + \mathcal{O}(\varepsilon\delta^2) \\ \dot{\varphi} &= \omega(b) + \mathcal{O}(\varepsilon).\end{aligned}$$

Similarly, the corresponding variational equation is of the form

$$\begin{pmatrix} \dot{B} \\ \dot{\Phi} \end{pmatrix} = \begin{pmatrix} \varepsilon A + \mathcal{O}(\varepsilon\delta) & \mathcal{O}(\varepsilon^3/\delta^2) \\ \mathcal{O}(1) & \mathcal{O}(\varepsilon^3/\delta^2) \end{pmatrix} \begin{pmatrix} B \\ \Phi \end{pmatrix}.$$

These relations and condition (4.3) imply that, for sufficiently small ε and for any fixed $\tau > 0$, the time- τ flow of (4.1) maps the strip $D = \{(b, \varphi) : \|b\| \leq \frac{1}{2}\delta, \varphi \in \mathbb{T}^d\}$ into itself, and the following bounds hold for the derivatives of the solution with respect to the initial values:

$$\begin{aligned}\left\| \frac{\partial b(\tau)}{\partial b(0)} \right\| &\leq L_{bb} = e^{-\tau\varepsilon\alpha} + \mathcal{O}(\varepsilon\delta), & \left\| \frac{\partial b(\tau)}{\partial \varphi(0)} \right\| &\leq L_{b\varphi} = \mathcal{O}(\varepsilon^3/\delta^2) \\ \left\| \frac{\partial \varphi(\tau)}{\partial b(0)} \right\| &\leq L_{\varphi b} = \mathcal{O}(1), & \left\| \frac{\partial \varphi(\tau)}{\partial \varphi(0)} - I \right\| &\leq L_{\varphi\varphi} = \mathcal{O}(\varepsilon^3/\delta^2).\end{aligned}\quad (4.5)$$

Hence, for sufficiently small ε ,

$$L_{\varphi\varphi} + L_{bb} + 2\sqrt{L_{\varphi b}L_{b\varphi}} \leq e^{-\tau\varepsilon\alpha/2} < 1.$$

Theorem 3.1 (and Exercise 1) used with φ, b in the roles of x, y now shows that the time- τ flow has an attractive invariant torus $\{(s(\varphi), \varphi) : \varphi \in \mathbb{T}^d\}$, where $s : \mathbb{T}^d \rightarrow \{\|b\| \leq \frac{1}{2}\delta\}$ is Lipschitz bounded by $\lambda = 2L_{b\varphi}/(1 - L_{\varphi\varphi} - L_{bb}) = \mathcal{O}(\varepsilon^3/\delta^2)$. This invariant torus attracts orbits of the time- τ flow map in the strip D with the attractivity factor $\lambda L_{\varphi b} + L_{bb} \leq e^{-\tau\varepsilon\alpha/2}$. As Exercise 2 shows, the torus is actually invariant for the differential equation (4.1). \square

XII.5 Weakly Attractive Invariant Tori of Numerical Integrators

Does the attractive invariant torus of Theorem 4.1 persist under numerical discretization of the perturbed integrable system? This question was first studied by Stoffer (1998) who worked directly with the discrete equations in his analysis. Here we take up the approach of Hairer & Lubich (1999) where the problem was studied by combining backward error analysis and perturbation theory, similar to what was done in the two preceding chapters.

XII.5.1 Modified Equations of Perturbed Differential Equations

Below we need to use backward error analysis for the numerical solution of a perturbed differential equation

$$\dot{y} = f(y) + \varepsilon g(y, \varepsilon), \quad y(0) = y_0 \quad (5.1)$$

with real-analytic functions f and g and small parameter ε . We consider applying a one-step method $y_1 = \Phi_h^\varepsilon(y_0)$ of order $p \geq 1$ with step size $h > 0$. The associated modified differential equations constructed in Chap. IX are then of the form

$$\dot{\tilde{y}} = \tilde{f}(\tilde{y}) + \varepsilon \tilde{g}(\tilde{y}, \varepsilon), \quad \tilde{y}(0) = y_0 \quad (5.2)$$

with suitably truncated series

$$\begin{aligned} \tilde{f}(y) &= f(y) + h^p f_{p+1}(y) + \dots + h^{N-1} f_N(y) \\ \tilde{g}(y, \varepsilon) &= g(y, \varepsilon) + h^p g_{p+1}(y, \varepsilon) + \dots + h^{N-1} g_N(y, \varepsilon), \end{aligned} \quad (5.3)$$

where the functions f_j are independent of ε, h, N , whereas the functions g_j are allowed to depend on ε . The following adapts Theorem IX.7.6 to the above situation.

Theorem 5.1. *Let $f(y) + \varepsilon g(y, \varepsilon)$ be real-analytic (in y and ε) and bounded by M for $y \in B_{2R}(y_0)$ and for all complex ε with $|\varepsilon| \leq \varepsilon_0$. Let the coefficients of the Taylor series (in h) of the numerical method be analytic in $B_R(y_0)$ with bounds (IX.7.5) for $|\varepsilon| \leq \varepsilon_0$. Then, there exists $h_0 > 0$ (proportional to R/M), such that for $h \leq h_0/4$ and for $N = N(h)$ the largest integer with $hN \leq h_0$, the difference between the numerical solution $y_1 = \Phi_h^\varepsilon(y_0)$ and the exact solution $\tilde{\varphi}_{N,t}^\varepsilon(y_0)$ of the truncated modified equation (5.2)-(5.3) satisfies*

$$\|\Phi_h^\varepsilon(y_0) - \tilde{\varphi}_{N,h}^\varepsilon(y_0)\| \leq Ch e^{-h_0/h}.$$

The functions \tilde{f} and \tilde{g} of (5.3) are real-analytic in $B_R(y_0)$ with

$$\|\tilde{f}(y) - f(y)\| \leq Ch^p, \quad \|\tilde{g}(y, \varepsilon) - g(y, \varepsilon)\| \leq Ch^p$$

for $y \in B_{R/2}(y_0)$ and $|\varepsilon| \leq \varepsilon_0$. The constants C are independent of $h \leq h_0/4$ and $|\varepsilon| \leq \varepsilon_0$.

Proof. The exponentially small estimate for $\Phi_h^\varepsilon(y_0) - \tilde{\varphi}_{N,h}^\varepsilon(y_0)$ is that of Theorem IX.7.6 applied to the differential equation (5.1). The $\mathcal{O}(h^p)$ bound for $\tilde{f}(y) - f(y)$ is the estimate (IX.7.14) applied to $\dot{y} = f(y)$. By applying that estimate to (5.1), a bound of the same type is obtained for $(\tilde{f}(y) + \varepsilon \tilde{g}(y, \varepsilon)) - (f(y) + \varepsilon g(y, \varepsilon))$, uniformly for all complex ε in the complex disk $|\varepsilon| \leq \varepsilon_0$. For any fixed $y \in B_{R/2}(y_0)$, the difference

$$\tilde{g}(y, \varepsilon) - g(y, \varepsilon) = \frac{1}{\varepsilon} \left([(\tilde{f}(y) + \varepsilon \tilde{g}(y, \varepsilon)) - (f(y) + \varepsilon g(y, \varepsilon))] - [\tilde{f}(y) - f(y)] \right)$$

is an analytic function of ε in the complex disk $|\varepsilon| \leq \varepsilon_0$, which is bounded by $\mathcal{O}(h^p)$ for $|\varepsilon| = \varepsilon_0$. By the maximum principle, the same bound then holds for $|\varepsilon| \leq \varepsilon_0$. \square

XII.5.2 Symplectic Methods

We apply a symplectic integrator with step size h to a real-analytic perturbed integrable Hamiltonian system in coordinates (p, q) ,

$$\begin{aligned}\dot{p} &= -\frac{\partial H}{\partial q}(p, q) + \varepsilon k(p, q) \\ \dot{q} &= \frac{\partial H}{\partial p}(p, q) + \varepsilon \ell(p, q).\end{aligned}\tag{5.4}$$

We assume that the unperturbed system ($\varepsilon = 0$) is a completely integrable system which satisfies the conditions of the Arnold–Liouville theorem, Theorem X.1.6. Hence, there exists a transformation to action-angle variables for the

integrable system: $(p, q) \mapsto (a, \theta)$ by Theorem X.1.6.

This change of coordinates transforms the integrable system to the equations $\dot{a} = 0$, $\dot{\theta} = \omega(a)$, and it transforms (5.4) to a system (4.1), for which we assume (4.2), (4.3) and the diophantine condition (X.2.4) with exponent ν for $\omega(a^*)$. The following theorem is a variant of results in Stoffer (1998) and Hairer & Lubich (1999). It shows that for symplectic methods, the invariant torus persists under a very mild restriction on the step size. For non-symplectic methods, this would require step sizes h with $h^p \ll \varepsilon$ (see Exercise 5).

Theorem 5.2. *Let a symplectic numerical integrator of order p be applied to a perturbed integrable Hamiltonian system (5.4) which satisfies the conditions stated above. Then, there exist $\varepsilon_0 > 0$ and $c_0 > 0$ such that, for $0 < \varepsilon \leq \varepsilon_0$ and for step sizes $h > 0$ satisfying*

$$h^p \leq c_0 |\log \varepsilon|^{-\kappa}\tag{5.5}$$

with $\kappa = \max(\nu + d + 1, p)$, the numerical method has an attractive invariant torus $\mathcal{T}_{\varepsilon, h}$. This torus is $\mathcal{O}(h^p)$ close to the invariant torus \mathcal{T}_ε of (5.4). It attracts an $\mathcal{O}(|\log \varepsilon|^{-2\kappa})$ neighbourhood with an exponential rate proportional to ε , uniformly in h .

Remark 5.3. The exponent $\nu + d + 1$ comes from Lemma X.4.1. It could be reduced to $\nu + 1$ by using Rüssmann’s estimates in place of that lemma; cf. the remark after Lemma X.4.1.

Proof of Theorem 5.2. The proof combines backward error analysis (Theorem IX.3.1 and Theorem 5.1), perturbation theory (Theorem X.4.4 and Lemma 2.1), and the invariant manifold theorem (Theorem 3.1).

(a) We begin by considering the symplectic method applied to the integrable Hamiltonian system (5.4) with $\varepsilon = 0$. This leads us back to the questions of Chap. X. We use backward error analysis and recall (Theorem IX.3.1) that the modified equation is again Hamiltonian and an $\mathcal{O}(h^p)$ perturbation of the integrable system, both in the (p, q) and the (a, θ) variables. We transform variables for the

modified equation of the integrable system: $(a, \theta) \mapsto (\tilde{a}, \tilde{\theta})$ by Theorem X.4.4,

with h^p in the role of the perturbation parameter. By (X.4.1) with N proportional to $|\log \varepsilon|$, and by condition (5.5) with a sufficiently small c_0 , the modified equations in these variables become

$$\begin{aligned}\dot{\tilde{a}} &= \mathcal{O}(\varepsilon^3) \\ \dot{\tilde{\theta}} &= \tilde{\omega}(\tilde{a}) + \mathcal{O}(\varepsilon^3)\end{aligned}\quad \text{for } \|\tilde{a} - a^*\| \leq c^* |\log \varepsilon|^{-2\kappa},$$

with $\tilde{\omega}(\tilde{a}) = \omega(\tilde{a}) + \mathcal{O}(h^p)$. Moreover, the transformation $(a, \theta) \mapsto (\tilde{a}, \tilde{\theta})$ is $\mathcal{O}(h^p)$ close to the identity.

(b) The modified equations of the perturbed system, written in the $(\tilde{a}, \tilde{\theta})$ variables, become

$$\begin{aligned}\dot{\tilde{a}} &= \varepsilon \tilde{r}(\tilde{a}, \tilde{\theta}) + \mathcal{O}(\varepsilon^3) \\ \dot{\tilde{\theta}} &= \tilde{\omega}(\tilde{a}) + \varepsilon \tilde{\rho}(\tilde{a}, \tilde{\theta}) + \mathcal{O}(\varepsilon^3)\end{aligned}\quad \text{for } \|\tilde{a} - a^*\| \leq c^* |\log \varepsilon|^{-2\kappa}, \quad (5.6)$$

where $\tilde{r}(\tilde{a}, \tilde{\theta}) = r(\tilde{a}, \tilde{\theta}) + \mathcal{O}(h^p)$ and $\tilde{\rho}(\tilde{a}, \tilde{\theta}) = \rho(\tilde{a}, \tilde{\theta}) + \mathcal{O}(h^p)$ by Theorem 5.1. Consider now these equations with the $\mathcal{O}(\varepsilon^3)$ terms dropped. We change variables for the

modified equation of the perturbed system: $(\tilde{a}, \tilde{\theta}) \mapsto (\tilde{b}, \tilde{\varphi})$ by Lemma 2.1.

(Note Exercise 4 with $\tilde{\omega}(a^*) = \omega(a^*) + \mathcal{O}(h^p)$ and (5.5).) The system (5.6) is transformed to the form of (4.4),

$$\begin{aligned}\dot{\tilde{b}} &= \varepsilon \tilde{m}(\tilde{b}) + \mathcal{O}(\varepsilon^3/\delta^2) \\ \dot{\tilde{\varphi}} &= \tilde{\omega}(\tilde{b}) + \varepsilon \tilde{\mu}(\tilde{b}) + \mathcal{O}(\varepsilon^3/\delta^2)\end{aligned}\quad (5.7)$$

with $\delta = c^* |\log \varepsilon|^{-2\kappa}$, and where $\tilde{m}(\tilde{b}) = \tilde{r}(\tilde{b}) + \mathcal{O}(\varepsilon/\delta) = \bar{r}(\tilde{b}) + \mathcal{O}(h^p) + \mathcal{O}(\varepsilon/\delta)$, and also the Jacobian of \tilde{m} at a^* is close to that of \bar{r} , so that it satisfies again (4.3), at least with α replaced by $\alpha/2$. In the same way as in the proof of Theorem 4.1 and with the same Lipschitz constants as in (4.5), we now obtain an attractive invariant torus of the modified equation of the perturbed system. The time- h flow of this equation is an exponentially small (in $1/h$) Lipschitz perturbation of the numerical one-step map, so that under condition (5.5) it is an $\mathcal{O}(\varepsilon^3)$ perturbation. Therefore, Theorem 3.1 yields an invariant torus $\mathcal{T}_{\varepsilon, h}$ of the numerical method.

(c) It remains to bound the distance between the tori $\mathcal{T}_{\varepsilon, h}$ and \mathcal{T}_ε . We recall that \mathcal{T}_ε was obtained by a transformation of the

perturbed system: $(a, \theta) \mapsto (b, \varphi)$ by Lemma 2.1,

which puts (4.1) into the form (4.4). We thus have the transformations

$$\begin{array}{ccc}
(a, \theta) & \xrightarrow{\varepsilon} & (b, \varphi) \\
h^p \downarrow & & \\
(\tilde{a}, \tilde{\theta}) & \xrightarrow{\varepsilon} & (\tilde{b}, \tilde{\varphi})
\end{array}$$

where the symbols h^p and ε indicate that the transformation is $\mathcal{O}(h^p)$ or $\mathcal{O}(\varepsilon)$ close to the identity. By the construction of Lemma 2.1, the composed transformation $(b, \varphi) \mapsto (\tilde{b}, \tilde{\varphi})$ is $\mathcal{O}(h^p)$ close to the identity and moreover, the right-hand sides of (4.4) and (5.7) differ by $\mathcal{O}(\varepsilon h^p)$. Theorem 3.2 (with $\rho = e^{-\varepsilon\tau\alpha/2}$) now shows that the functions $s_{\varepsilon, h}$ and s_ε defining $\mathcal{T}_{\varepsilon, h}$ and \mathcal{T}_ε , respectively, differ by $\mathcal{O}(h^p)$. This yields the desired distance bound. \square

XII.5.3 Symmetric Methods

A result analogous to the theorem of the previous subsection holds for reversible methods applied to perturbed reversible systems

$$\begin{aligned}
\dot{u} &= f(u, v) + \varepsilon k(u, v) \\
\dot{v} &= g(u, v) + \varepsilon \ell(u, v)
\end{aligned}$$

where the unperturbed system ($\varepsilon = 0$) is a real-analytic integrable reversible system. If the perturbed system, written in action-angle variables of the unperturbed system, satisfies the conditions of Theorem 4.1, then a reversible analogue of Theorem 5.2 holds, where the terms “symplectic” and “Hamiltonian” are simply replaced by “reversible”. The proof remains the same, working with the reversible analogues of the results used for the Hamiltonian case.

XII.6 Exercises

1. In the situation of the invariant manifold theorem, Theorem 3.1, suppose in addition that f and g are α -periodic in x : $f(x + \alpha, y) = f(x, y)$, $g(x + \alpha, y) = g(x, y)$ for all $x \in X$, $y \in Y$. Show that in this case the function s defining the invariant manifold is also α -periodic.

Hint. The Hadamard transform maps α -periodic functions to α -periodic functions.

2. Show that if the time- τ flow map $\Phi = \varphi_\tau$ of a differential equation has an attractive invariant manifold \mathcal{M} , and if the flow φ_t maps a domain of attractivity of \mathcal{M} under Φ into itself for every real t , then \mathcal{M} is also invariant under the flow φ_t for every real t .

Hint. Write $\varphi_t = \Phi^n \circ \varphi_t \circ \Phi^{-n}$ and use the attractivity of \mathcal{M} for $n \rightarrow \infty$.

3. Prove that in the situation of Theorem 3.1, iterates $(x_{n+1}, y_{n+1}) = \Phi(x_n, y_n)$ have the *property of asymptotic phase* (Nipp & Stoffer 1992): there exists a sequence $(\tilde{x}_n, \tilde{y}_n)$ of iterates on the invariant manifold, i.e., with $(\tilde{x}_{n+1}, \tilde{y}_{n+1}) = \Phi(\tilde{x}_n, \tilde{y}_n)$ and $\tilde{y}_n = s(\tilde{x}_n)$, such that for all $n \geq 0$,

$$\begin{aligned} \|x_n - \tilde{x}_n\| &\leq c \|y_n - s(x_n)\| \\ \|y_n - \tilde{y}_n\| &\leq (1 + \lambda c) \|y_n - s(x_n)\|, \end{aligned}$$

where $c = \lambda/(1 - \lambda\lambda^*)$ with $\lambda = 2L_{yx}/(1 - L_{xx} - L_{yy})$ of (3.5) and $\lambda^* = 2L_{xy}/(1 - L_{xx} - L_{yy})$. Note that $\|y_n - s(x_n)\| \leq \rho^n \|y_0 - s(x_0)\|$ by Theorem 3.1.

Hint. Consider the sequences $(\tilde{x}_n^{(k)}, \tilde{y}_n^{(k)})$ defined by $\tilde{x}_k^{(k)} = x_k, \tilde{y}_k^{(k)} = s(x_k)$ and $(\tilde{x}_{n+1}^{(k)}, \tilde{y}_{n+1}^{(k)}) = \Phi(\tilde{x}_n^{(k)}, \tilde{y}_n^{(k)})$ for $n = k-1, \dots, 1, 0$. Show that, for fixed n , the sequence $(x_n^{(k)})$ ($k \geq n$) is a Cauchy sequence.

4. Show that Lemma 2.1 holds unchanged if the diophantine condition (X.2.4) for $\omega(a^*)$ is weakened to $\omega(a^*) = \omega^* + \mathcal{O}(\delta^2)$ with ω^* satisfying (X.2.4).
5. In the situation of Theorem 5.2, show that every numerical integrator of order p has an attractive invariant torus if $h^p \ll \varepsilon$. This torus is $\mathcal{O}(h^p/\varepsilon)$ close to the invariant torus of the continuous system.

Chapter XIII.

Oscillatory Differential Equations with Constant High Frequencies

This chapter deals with numerical methods for second-order differential equations with oscillatory solutions. These methods are designed to require a new complete function evaluation only after a time step over one or many periods of the fastest oscillations in the system. Various such methods have been proposed in the literature – some of them decades ago, some very recently, motivated by problems from molecular dynamics, astrophysics and nonlinear wave equations. For these methods it is not obvious what implications geometric properties like symplecticity or reversibility have on the long-time behaviour, e.g., on energy conservation. The backward error analysis of Chap. IX, which was the backbone of the results of the three preceding chapters, is no longer applicable when the product of the step size with the highest frequency is not small, which is the situation of interest here. The “exponentially small” remainder terms are now only $\mathcal{O}(1)$! For differential equations where the high frequencies of the oscillations remain nearly constant along the solution, a substitute for the backward error analysis of Chap. IX is given by the *modulated Fourier expansions* of the exact and the numerical solutions. Among other properties, they permit us to understand the numerical long-time conservation of the total and oscillatory energies (or the failure of conserving energy in certain cases). It turns out, symmetry of the methods is still essential, but symplecticity plays no role in the analysis and in the numerical experiments, and new conditions of an apparently non-geometric nature come into play.

XIII.1 Towards Longer Time Steps in Solving Oscillatory Equations of Motion

Dynamical systems with multiple time scales pose a major problem in simulations because the small time steps required for stable integration of the fast motions lead to large numbers of time steps required for the observation of slow degrees of freedom and thus to the need to compute a large number of forces.

(M. Tuckerman, B.J. Berne & G.J. Martyna 1992)

We describe numerical methods that have been proposed for solving highly oscillatory second-order differential equations with fewer force evaluations than are needed by standard integrators like the Störmer–Verlet method. We present the ideas

underlying the construction of the methods and leave numerical comparisons to Sect. XIII.2 and the analysis of the methods to Sections XIII.3–XIII.6. We consider only methods that are symmetric or symplectic. The presentation in this section follows roughly the chronological order.

XIII.1.1 The Störmer–Verlet Method vs. Multiple Time Scales

Perhaps the most widely used method of integrating the equations of motion is that initially adopted by Verlet (1967) and attributed to Störmer.
(M.P. Allen & D.J. Tildesley 1987, p. 78)

The Newtonian equations of motion of particle systems (in molecular dynamics, astrophysics and elsewhere) are second-order differential equations

$$\ddot{q} = -\nabla V(q). \quad (1.1)$$

To simplify the presentation, we omit the positive definite mass matrix M which would usually multiply \ddot{q} . This entails no loss of generality, since a transformation $q \rightarrow M^{1/2}q$ and $V(q) \rightarrow V(M^{-1/2}q)$ gives the very form (1.1).

The standard numerical integrator of molecular dynamics is the Störmer–Verlet scheme; see Chap. I. We recall that this method computes the new positions q_{n+1} at time t_{n+1} from

$$q_{n+1} - 2q_n + q_{n-1} = h^2 f_n \quad (1.2)$$

with the force $f_n = -\nabla V(q_n)$. Velocity approximations are given by

$$\dot{q}_n = \frac{q_{n+1} - q_{n-1}}{2h}.$$

In its one-step formulation (see (I.1.17)) the method reads¹

$$\begin{aligned} p_{n+1/2} &= p_n + \frac{1}{2}hf_n \\ q_{n+1} &= q_n + hp_{n+1/2} \\ p_{n+1} &= p_{n+1/2} + \frac{1}{2}hf_{n+1}. \end{aligned} \quad (1.3)$$

We recall that this is a symmetric and symplectic method of order 2. For linear stability, i.e., for bounded error propagation in linearized equations, the step size must be restricted to

$$h\omega < 2$$

where ω is the largest eigenfrequency (i.e., square root of an eigenvalue) of the Hessian matrix $\nabla^2 V(q)$ along the numerical solution; see Sect. I.5.1. Good energy conservation requires an even stronger restriction on the step size. Values of $h\omega \approx \frac{1}{2}$ are frequently used in molecular dynamics simulations.

The potential $V(q)$ is often a sum of potentials that act on different time scales,

¹ We write p when the Hamiltonian structure and symplecticity are an issue, and \dot{q} otherwise.

$$V(q) = W(q) + U(q) \quad \text{with} \quad \nabla^2 W(q) \text{ positive semi-definite and} \quad (1.4)$$

$$\|\nabla^2 W(q)\| \gg \|\nabla^2 U(q)\| .$$

In this situation, solutions are in general highly oscillatory on the slow time scale $\tau \sim 1/\|\nabla^2 U(q)\|^{1/2}$.

In particular when the *fast* forces $-\nabla W(q)$ are cheaper to evaluate than the *slow* forces $-\nabla U(q)$, it is of interest to devise methods where the required number of slow-force evaluations is not (or not severely) affected by the presence of the fast forces which are responsible for the oscillatory behaviour and which restrict the step size of standard integrators like the Störmer–Verlet scheme. This situation occurs in molecular dynamics, where $W(q)$ corresponds to short-range molecular bonds, whereas $U(q)$ includes *inter alia* long-range electrostatic potentials.

In some approaches to this computational problem, the differential model is modified: highly oscillatory components are replaced by constraints (Ryckaert, Cicotti & Berendsen 1977), or stochastic and dissipative terms are added to the model (see Schlick 1999). Such modifications may prove highly successful in some applications. In the following, however, we restrict our attention to methods which aim at long time steps directly for the problem (1.1) with (1.4).

Spatial semi-discretizations of nonlinear wave equations, such as the sine-Gordon equation

$$u_{tt} = u_{xx} - \sin u ,$$

form another important class of equations (1.1) with (1.4). Here $W(q) = \frac{1}{2}q^T A q$, where A is the discretization matrix of the differential operator $-\partial^2/\partial x^2$.

XIII.1.2 Gautschi's and Deuffhard's Trigonometric Methods

It is anticipated that trigonometric methods can be applied, with similar success, also to nonlinear differential equations describing oscillation phenomena. (W. Gautschi 1961)

The oldest methods allowing the use of long time steps in oscillatory problems concern the particular case of a quadratic potential $W(q) = \frac{1}{2}\omega^2 q^T q$ with $\omega \gg 1$, for which the equations take the form

$$\ddot{q} = -\omega^2 q + g(q) . \quad (1.5)$$

For such equations, Gautschi (1961) proposed a number of methods of multistep type which are constructed to be exact if the solution is a trigonometric polynomial in ωt of a prescribed degree. The simplest of these methods (and the only symmetric one) reads

$$q_{n+1} - 2q_n + q_{n-1} = h^2 \operatorname{sinc}^2(\tfrac{1}{2}h\omega) \ddot{q}_n , \quad (1.6)$$

where $\operatorname{sinc} \xi = \sin \xi / \xi$ and $\ddot{q}_n = -\omega^2 q_n + g_n$ with $g_n = g(q_n)$, or equivalently

$$q_{n+1} - 2 \cos(h\omega) q_n + q_{n-1} = h^2 \operatorname{sinc}^2(\tfrac{1}{2}h\omega) g_n . \quad (1.7)$$

The method gives the exact solution for equations (1.5) with $g = \text{Const}$ and arbitrary ω (see also Hersch (1958) for such a construction principle). This property is readily verified with the variation-of-constants formula

$$\begin{pmatrix} q(t) \\ \dot{q}(t) \end{pmatrix} = \begin{pmatrix} \cos t\omega & \omega^{-1} \sin t\omega \\ -\omega \sin t\omega & \cos t\omega \end{pmatrix} \begin{pmatrix} q_0 \\ \dot{q}_0 \end{pmatrix} + \int_0^t \begin{pmatrix} \omega^{-1} \sin(t-s)\omega \\ \cos(t-s)\omega \end{pmatrix} g(q(s)) ds. \quad (1.8)$$

This formula also shows that the following scheme for a velocity approximation becomes exact for $g = \text{Const}$:

$$\dot{q}_{n+1} - \dot{q}_{n-1} = 2h \operatorname{sinc}(h\omega) \ddot{q}_n. \quad (1.9)$$

Starting values q_1 and \dot{q}_1 are also obtained from (1.8) with $g(q_0)$ in place of $g(q(s))$.

Deuffhard (1979) considered h^2 -extrapolation based on the explicit symmetric method that is obtained by replacing the integral term in (1.8) by its trapezoidal rule approximation:

$$\begin{pmatrix} q_{n+1} \\ h\dot{q}_{n+1} \end{pmatrix} = \begin{pmatrix} \cos h\omega & \operatorname{sinc} h\omega \\ -h\omega \sin h\omega & \cos h\omega \end{pmatrix} \begin{pmatrix} q_n \\ h\dot{q}_n \end{pmatrix} + \frac{h^2}{2} \begin{pmatrix} \operatorname{sinc}(h\omega) g_n \\ g_{n+1} + \cos(h\omega) g_n \end{pmatrix}. \quad (1.10)$$

Eliminating the velocities yields the two-step formulation

$$q_{n+1} - 2 \cos(h\omega) q_n + q_{n-1} = h^2 \operatorname{sinc}(h\omega) g_n. \quad (1.11)$$

The velocity approximation is obtained back from

$$2h \operatorname{sinc}(h\omega) \dot{q}_n = q_{n+1} - q_{n-1} \quad (1.12)$$

or alternatively from

$$\dot{q}_{n+1} - 2 \cos(h\omega) \dot{q}_n + \dot{q}_{n-1} = h^2 \frac{g_{n+1} - g_{n-1}}{2h}.$$

Both Gautschi's and Deuffhard's method reduce to the Störmer–Verlet scheme for $\omega = 0$. Both methods extend in a straightforward way to systems

$$\ddot{q} = -Aq + g(q) \quad (1.13)$$

with a symmetric positive semi-definite matrix A , by formally replacing ω by $\Omega = A^{1/2}$ in the above formulas. The methods then require the computation of products of entire functions of the matrix $h^2 A$ with vectors. This can be done by diagonalizing A , which is efficient for problems of small dimension or in spectral methods for nonlinear wave equations. In high-dimensional problems where a diagonalization is not feasible, these matrix function times vector products can be efficiently computed by superlinearly convergent Krylov subspace methods, see Druskin & Knizhnerman (1995) and Hochbruck & Lubich (1997).

The above methods permit extensions to more general problems (1.1) with (1.4), but this requires a reinterpretation to which we turn next.

XIII.1.3 The Impulse Method

Integrators based on r-RESPA [...] have led to considerable speed-up in the CPU time for large scale simulations of biomacromolecular solutions. Since r-RESPA is symplectic such integrators are very stable.

(B.J. Berne 1999)

The Störmer–Verlet method (1.3) can be interpreted as approximating the flow φ_h^H of the system with Hamiltonian $H(p, q) = T(p) + V(q)$ with $T(p) = \frac{1}{2}p^T p$ by the symmetric splitting

$$\varphi_{h/2}^V \circ \varphi_h^T \circ \varphi_{h/2}^V,$$

which involves only the flows of the systems with Hamiltonians $T(p)$ and $V(q)$, which are trivial to compute; see Sect. II.5.

In the situation (1.4) of a potential $V = W + U$, we may instead use a different splitting of $H = (T + W) + U$ and approximate the flow φ_h^H of the system by

$$\varphi_{h/2}^U \circ \varphi_h^{T+W} \circ \varphi_{h/2}^U.$$

This gives a method that was proposed in the context of molecular dynamics by Grubmüller, Heller, Windemuth & Schulten (1991) (their Verlet-I scheme) and by Tuckerman, Berne & Martyna (1992) (their r-RESPA scheme). Following the terminology of García-Archilla, Sanz-Serna & Skeel (1999) we here refer to this method as the *impulse method*:

1. kick: set $p_n^+ = p_n - \frac{1}{2}h \nabla U(q_n)$
 2. oscillate: solve $\ddot{q} = -\nabla W(q)$ with initial values (q_n, p_n^+)
over a time step h to obtain (q_{n+1}, p_{n+1}^-)
 3. kick: set $p_{n+1} = p_{n+1}^- - \frac{1}{2}h \nabla U(q_{n+1})$
- (1.14)

Step 2 must in general be computed approximately by a numerical integrator with a smaller time step, which results in the multiple time stepping method that we encountered in Sect. VIII.4. If the inner integrator is symplectic and symmetric, as it would be for the natural choice of the Störmer–Verlet method, then also the overall method is symplectic – as a composition of symplectic transformations, and it is symmetric – as a symmetric composition of symmetric steps.

It is interesting to note that the impulse method (with exact solution of step 2) reduces to Deuffhard’s method in the case of a quadratic potential $W(q) = \frac{1}{2}q^T Aq$ (Exercise 1).

Though the method does allow larger step sizes than the Störmer–Verlet method in molecular dynamics simulations, it is not free from numerical difficulties. Biesadecki & Skeel (1993) and García-Archilla et al. (1999) report and in linear model problems analyze instabilities and numerical resonance phenomena when the product of the step size h with an eigenfrequency ω of $\nabla^2 W$ is near an integral multiple of π .

XIII.1.4 The Mollified Impulse Method

We also propose a nontrivial improvement of the impulse method that we call the *mollified impulse method*, for which superior stability and accuracy is demonstrated.

(B.García-Archilla, J.M. Sanz-Serna & R.D. Skeel 1999)

Difficulties with the impulse method can be intuitively seen to come from two sources: the slow force $-\nabla U(q)$ has an effect only at the ends of a time step, but it does not enter into the oscillations in between; the slow force is evaluated, somewhat arbitrarily, at isolated points of the oscillatory solution.

García-Archilla et al. (1999) propose to evaluate the slow force at an *averaged* value $\bar{q}_n = a(q_n)$. They replace the potential $U(q)$ by $\bar{U}(q) = U(a(q))$ and hence the slow force $-\nabla U(q)$ in the impulse method by the *mollified force*

$$-\nabla \bar{U}(q) = -a'(q)^T \nabla U(a(q)) . \quad (1.15)$$

Since this *mollified impulse method* is the impulse method for a modified potential, it is again symplectic and symmetric.

There are numerous possibilities to choose the average $a(q_n)$, but care should be taken that it is only a function of the position q_n and thus independent of p_n , in order to obtain a symplectic and symmetric method. This precludes taking averages of the solution of the problem in the oscillation step (Step 2) of the algorithm. Instead, one solves the auxiliary initial value problem

$$\ddot{x} = -\nabla W(x) \quad \text{with} \quad x(0) = q, \quad \dot{x}(0) = 0 \quad (1.16)$$

together with the variational equation (using the same method and the same step size)

$$\ddot{X} = -\nabla^2 W(x(t))X \quad \text{with} \quad X(0) = I, \quad \dot{X}(0) = 0 \quad (1.17)$$

and computes the time average over an interval of length ch for some $c > 0$:

$$a(q) = \frac{1}{ch} \int_0^{ch} x(t) dt, \quad a'(q) = \frac{1}{ch} \int_0^{ch} X(t) dt . \quad (1.18)$$

García-Archilla et al. (1999) found that the choice $c = 1$ gives the best results. Weighted averages instead of the simple average used above give no improvement.

Izaguirre, Reich & Skeel (1999) propose to take $a(q)$ as a projection of q to the manifold $\nabla W(q) = 0$ of rest positions of the fast forces, for situations where all non-zero eigenfrequencies of $\nabla^2 W(q)$ are much larger than those of $\nabla^2 U(q)$. This choice is motivated by the fact that solutions oscillate about this manifold.

We now turn to the interesting special case of a quadratic $W(q) = \frac{1}{2}q^T Aq$ with a symmetric positive semi-definite matrix A . In this case, the above average can be computed analytically. It becomes

$$a(q) = \phi(h\Omega)q$$

with $\Omega = A^{1/2}$ and the function $\phi(\xi) = \text{sinc}(c\xi)$. For $a(q)$ defined by the orthogonal projection to $Aq = 0$ we have $\phi(0) = 1$ and $\phi(\xi) = 0$ for ξ away from 0. With $g_n = -\phi(h\Omega)\nabla U(\phi(h\Omega)q_n)$, the mollified impulse method reduces to

$$\begin{aligned} p_n^+ &= p_n + \frac{1}{2}hg_n \\ \begin{pmatrix} q_{n+1} \\ p_{n+1}^- \end{pmatrix} &= \begin{pmatrix} \cos h\Omega & h \text{sinc } h\Omega \\ -\Omega \sin h\Omega & \cos h\Omega \end{pmatrix} \begin{pmatrix} q_n \\ p_n^+ \end{pmatrix} \\ p_{n+1} &= p_{n+1}^- + \frac{1}{2}hg_{n+1}. \end{aligned} \quad (1.19)$$

This can equivalently be written as (1.10) with the same g_n (and Ω in place of ω), or in the two-step form (1.11) with (1.12).

XIII.1.5 Gautschi's Method Revisited

We recall that Gautschi's method (1.7) (with $\Omega = A^{1/2}$ in place of ω) integrates equations $\ddot{q} = -Aq + g(q)$ exactly in the case of a constant inhomogeneity $g(q) = \text{Const}$. This property is obviously kept if the argument of g in the algorithm is modified to

$$g_n = g(\phi(h\Omega)q_n)$$

similar to the previous subsection. Such Gautschi-type methods were analyzed by Hochbruck & Lubich (1999a). Functions ϕ with $\phi(0) = 1$ that vanish at integral multiples of π give a substantial improvement over the original Gautschi method. The choice

$$\phi(\xi) = \text{sinc } \xi \left(1 + \frac{1}{3} \sin^2 \frac{1}{2}\xi\right) \quad (1.20)$$

was found to give particularly good accuracy. The methods are symmetric but not symplectic.

The following symmetric method for general problems (1.1) with (1.4) was proposed by Hochbruck & Lubich (1999a). The method reduces to Gautschi-type methods for quadratic $W(q) = \frac{1}{2}q^T Aq$. Given q_n and \dot{q}_n , one computes an averaged value $\bar{q}_n = a(q_n)$ and the solution of

$$\ddot{u} = -\nabla W(u) - \nabla U(\bar{q}_n) \quad \text{with} \quad u(0) = q_n, \quad \dot{u}(0) = \dot{q}_n \quad (1.21)$$

backwards and forwards on the intervals from 0 to $-h$ and 0 to h . Note that this requires only one evaluation of the slow force $-\nabla U$. Then, q_{n+1} and \dot{q}_{n+1} are computed from

$$\begin{aligned} q_{n+1} - 2q_n + q_{n-1} &= u(h) - 2u(0) + u(-h) \\ \dot{q}_{n+1} - \dot{q}_{n-1} &= \dot{u}(h) - \dot{u}(-h). \end{aligned} \quad (1.22)$$

When the differential equation for u is solved approximately by a symmetric numerical method with smaller time steps, then this becomes a symmetric multiple time-stepping method. For the interpretation as an averaged-force method and for the corresponding one-step version, where the initial value for the velocity in (1.21) is replaced by $\dot{u}(0) = 0$, we refer back to Sect. VIII.4 (where q_n instead of the average $\bar{q}_n = a(q_n)$ was taken as the argument of the slow force $-\nabla U$).

XIII.1.6 Two-Force Methods

Hairer & Lubich (2000a) compare the analytical solution and the numerical solutions given by the above methods in the Fermi–Pasta–Ulam model of Sect. I.5.1, using the tool of modulated Fourier expansions (see Sections XIII.3 and XIII.5 below). Their analysis of the slow energy exchange between stiff springs leads them to propose the following method for equations $\ddot{q} = -Aq + g(q)$, which requires two evaluations of the slow force per time step: with $\Omega = A^{1/2}$, set

$$q_{n+1} - 2 \cos(h\Omega) q_n + q_{n-1} = h^2 \operatorname{sinc}(h\Omega) g(q_n) + h^2 d_n \quad (1.23)$$

with

$$d_n = \operatorname{sinc}^2(h\Omega) g(q_n) - \operatorname{sinc}(h\Omega) g(\operatorname{sinc}(h\Omega) q_n). \quad (1.24)$$

This method gives the correct slow energy exchange between stiff components in the model problem and has better energy conservation than the Deuffhard/impulse method. With the velocity approximation (1.12) the method can equivalently be written in the one-step forms (1.19) or (1.10). The method extends again to a symmetric method for general problems (1.1) with (1.4), giving a correction to the impulse method: let $g(q) = -\nabla U(q)$ and let $a(q)$ be defined by (1.18) with $c = 1$. Set $\bar{q}_n = a(q_n)$ and

$$\bar{g}(q_n) = \frac{2}{h^2} \left(a\left(q_n + \frac{1}{2} h^2 g(q_n)\right) - a(q_n) \right).$$

The method then consists of taking

$$g_n = g(q_n) + \bar{g}(q_n) - g(\bar{q}_n)$$

instead of $g(q_n) = -\nabla U(q_n)$ in the impulse method (1.14).

A two-force method with interesting properties, for situations where all non-zero eigenfrequencies of A are much larger than those of $\nabla^2 U(q)$, is given by (1.23) with

$$d_n = \operatorname{sinc}^2(\tfrac{1}{2} h\Omega) g(\chi(h\Omega) q_n) - \operatorname{sinc}(h\Omega) g(\chi(h\Omega) q_n), \quad (1.25)$$

where $\chi(0) = 1$ and $\chi(\xi) = 0$ for ξ away from 0.

XIII.2 A Nonlinear Model Problem and Numerical Phenomena

To gain insight into the properties of the various numerical methods described in the previous section, it is helpful to study the methods when they are applied to suitably chosen, rather simple model problems which show characteristic features but are still accessible to an analysis. Such an approach has traditionally been very successful for stiff differential equations (see, e.g., Hairer & Wanner 1996). For the

present stiff-oscillatory case we investigate the behaviour of the numerical methods on nonlinear systems

$$\ddot{x} + \Omega^2 x = g(x) \quad (2.1)$$

with a smooth gradient nonlinearity $g(x) = -\nabla U(x)$ and with the square matrix

$$\Omega = \begin{pmatrix} 0 & 0 \\ 0 & \omega I \end{pmatrix}, \quad \omega \gg 1, \quad (2.2)$$

with blocks of arbitrary dimension. We consider only solutions whose energy is bounded independently of ω , so that in particular the initial values satisfy

$$\frac{1}{2} \|\dot{x}(0)\|^2 + \frac{1}{2} \|\Omega x(0)\|^2 \leq E \quad (2.3)$$

with E independent of ω .

The Fermi–Pasta–Ulam (FPU) problem of Sect. I.5.1 belongs precisely to this class, and we will present numerical experiments with this example. In the model problem (2.1) with (2.2) we clearly impose strong restrictions in that the high frequencies are confined to the linear part and that there is a single, constant high frequency. The extension to several high frequencies will be given in Sect. XIII.9, and constant-frequency systems with a position-dependent kinetic energy term are considered in Sect. XIII.10. Oscillatory systems with time- or solution-dependent high frequencies will be studied, with different techniques and for different numerical methods, in Chap. XIV.

In any case, satisfactory behaviour of a method on the model problem (2.1) can be anticipated to be necessary for a successful treatment of more general situations.

XIII.2.1 Time Scales in the Fermi–Pasta–Ulam Problem

The FPU model shows different behaviour on different time scales: almost-harmonic motion of the stiff springs on the time scale ω^{-1} , motion of the soft springs on the scale ω^0 , energy exchange between stiff springs on the time scale ω , and almost-preservation of the oscillatory energy over intervals that are exponentially long in ω . This is illustrated in the following.

We consider the FPU problem with three stiff springs with the data of Sect. I.5.1. The four pictures of Fig. 2.1 show the evolution of the following quantities: the total energy

$$H(x, \dot{x}) = \frac{1}{2} \dot{x}^T \dot{x} + \frac{1}{2} x^T \Omega^2 x + U(x), \quad (2.4)$$

(or rather $H - 0.8$ for graphical reasons), which is a conserved quantity; the oscillatory energy

$$I = I_1 + I_2 + I_3 \quad \text{with} \quad I_j = \frac{1}{2} \dot{x}_{1,j}^2 + \frac{1}{2} \omega^2 x_{1,j}^2, \quad (2.5)$$

where $x_{1,j}$ is the j th component of the lower half $x_1 \in \mathbb{R}^3$ of $x = (x_0, x_1)^T \in \mathbb{R}^6$, decomposed according to the blocks of Ω in (2.2). We recall that $x_{1,j}$ represents the

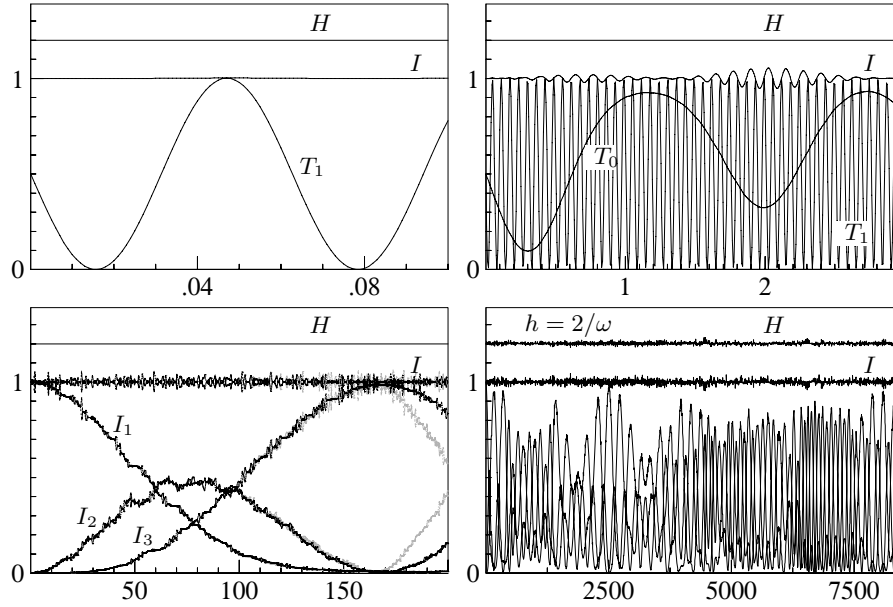


Fig. 2.1. Different time scales in the Fermi–Pasta–Ulam model ($\omega = 50$)

elongation of the j th stiff spring. Further quantities shown are the kinetic energy of the mass centre motion and of the relative motion of masses joined by a stiff spring,

$$T_0 = \frac{1}{2} \|\dot{x}_0\|^2, \quad T_1 = \frac{1}{2} \|\dot{x}_1\|^2.$$

Time Scale ω^{-1} . The vibration of the stiff linear springs is nearly harmonic with almost-period π/ω . This is illustrated by the plot of T_1 in the first picture.

Time Scale ω^0 . This is the time scale of the motion of the soft nonlinear springs, as is exemplified by the plot of T_0 in the second picture of Fig. 2.1.

Time Scale ω . A slow energy exchange among the stiff springs takes place on the scale ω . In the third picture, the initially excited first stiff spring passes energy to the second one, and then also the third stiff spring begins to vibrate. The picture also illustrates that the problem is very sensitive to perturbations of the initial data: the grey curves of each of I_1, I_2, I_3 correspond to initial data where 10^{-5} has been added to $x_{0,1}(0)$, $\dot{x}_{0,1}(0)$ and $\dot{x}_{1,1}(0)$. The displayed solutions of the first three pictures have been computed very accurately by an adaptive integrator.

Time Scale ω^N , $N \geq 2$. The oscillatory energy I has only $\mathcal{O}(\omega^{-1})$ deviations from the initial value over very long time intervals. The fourth picture of Fig. 2.1 shows the total energy H and the oscillatory energy I as computed by method (1.10)–(1.11) of Sect. XIII.1.2 with the step size $h = 2/\omega$, which is nearly as large as the length of the time interval of the first picture. No drift is seen for H or I .

XIII.2.2 Numerical Methods

The methods described in Sect. XIII.1 all have in common that they reduce to the Störmer–Verlet method when they are applied to (2.1) with $\Omega = 0$, and they become exact solvers for the linear homogeneous problem with $g(x) \equiv 0$. They can be formulated as one-step or two-step schemes.

Two-Step Formulation. All the methods of Sections XIII.1.2–XIII.1.5, when applied to the system (2.1), can be written in the two-step form

$$x_{n+1} - 2 \cos(h\Omega) x_n + x_{n-1} = h^2 \Psi g(\Phi x_n). \quad (2.6)$$

Here $\Psi = \psi(h\Omega)$ and $\Phi = \phi(h\Omega)$, where the *filter functions* ψ and ϕ are even, real-valued functions with $\psi(0) = \phi(0) = 1$. In our numerical experiments we will consider the following choices of ψ and ϕ , where again $\text{sinc}(\xi) = \sin \xi / \xi$:

(A)	$\psi(\xi) = \text{sinc}^2(\frac{1}{2}\xi)$	$\phi(\xi) = 1$	Gautschi (1961)
(B)	$\psi(\xi) = \text{sinc}(\xi)$	$\phi(\xi) = 1$	Deuffhard (1979)
(C)	$\psi(\xi) = \text{sinc}(\xi) \phi(\xi)$	$\phi(\xi) = \text{sinc}(\xi)$	García-Archilla & al. (1999)
(D)	$\psi(\xi) = \text{sinc}^2(\frac{1}{2}\xi)$	$\phi(\xi)$ of (1.20)	Hochbruck & Lubich (1999a)
(E)	$\psi(\xi) = \text{sinc}^2(\xi)$	$\phi(\xi) = 1$	Hairer & Lubich (2000a)

One-Step Formulation. The method (2.6) can be written as a symmetric one-step method of a form that is motivated by the variation-of-constants formula (1.8). This now also includes a velocity approximation \dot{x}_n :

$$x_{n+1} = \cos h\Omega x_n + \Omega^{-1} \sin h\Omega \dot{x}_n + \frac{1}{2} h^2 \Psi g_n \quad (2.7)$$

$$\dot{x}_{n+1} = -\Omega \sin h\Omega x_n + \cos h\Omega \dot{x}_n + \frac{1}{2} h (\Psi_0 g_n + \Psi_1 g_{n+1}) \quad (2.8)$$

where $g_n = g(\Phi x_n)$ and $\Psi_0 = \psi_0(h\Omega)$, $\Psi_1 = \psi_1(h\Omega)$ with even functions ψ_0 , ψ_1 satisfying $\psi_0(0) = 1$, $\psi_1(0) = 1$. Exchanging $n \leftrightarrow n+1$ and $h \leftrightarrow -h$ in the method, it is seen that the method is symmetric if and only if

$$\psi(\xi) = \text{sinc}(\xi) \psi_1(\xi), \quad \psi_0(\xi) = \cos(\xi) \psi_1(\xi). \quad (2.9)$$

The method is then symplectic if and only if (Exercise 2)

$$\psi(\xi) = \text{sinc}(\xi) \phi(\xi). \quad (2.10)$$

Two-Step Velocity Schemes. For a symmetric method (2.7)–(2.8) the velocity approximation can be equivalently obtained from

$$2h \text{sinc}(h\omega) \dot{x}_n = x_{n+1} - x_{n-1} \quad (2.11)$$

(for $\sin(h\omega) \neq 0$) or from

$$\dot{x}_{n+1} - 2 \cos(h\Omega) \dot{x}_n + \dot{x}_{n-1} = \frac{1}{2} h \Psi_1 (g_{n+1} - g_{n-1}). \quad (2.12)$$

The latter formula gives a symmetric two-step method for arbitrary even functions ψ_1 with $\psi_1(0) = 1$, which do not necessarily satisfy (2.9).

Multi-Force Methods. The methods of Sect. XIII.1.6 belong to the class of multi-force methods, which generalize the right-hand side of (2.6) to a linear combination of such terms:

$$x_{n+1} - 2\cos(h\Omega)x_n + x_{n-1} = h^2 \sum_{j=1}^k \Psi_j g(\Phi_j x_n) \quad (2.13)$$

with $\Psi_j = \psi_j(h\Omega)$, $\Phi_j = \phi_j(h\Omega)$, where ψ_j, ϕ_j are even functions with

$$\sum_{j=1}^k \psi_j(0) = 1, \quad \phi_j(0) = 1 \quad \text{for } j = 1, \dots, k.$$

In our numerical experiments we include the method

$$(F) \quad \text{two-force method (1.23) with (1.24).}$$

XIII.2.3 Accuracy Comparisons

The accuracy of the methods (A)-(E) and the Störmer-Verlet method on a short time interval is shown in Fig. 2.2, where the errors at $t = 1$ of the different solution components in the FPU problem (with $\omega = 50$) are plotted as a function of the step size h . Here and in all the following numerical experiments, the methods were

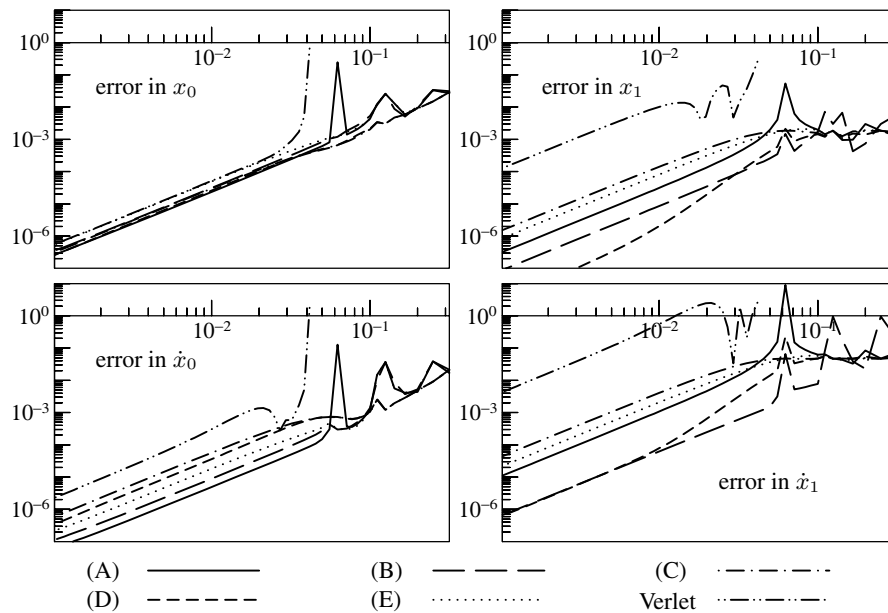


Fig. 2.2. Global error at $t = 1$ for the different components and for the five methods (A) - (E) and the Störmer-Verlet method as a function of the step size h

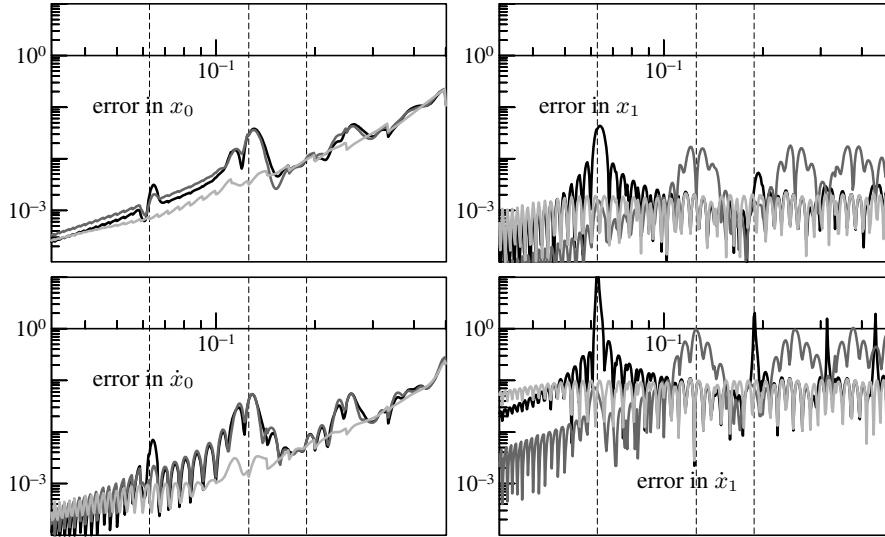


Fig. 2.3. Global error at the first grid point after $t = 1$ for the different components as a function of the step size h . The error for method (A) is drawn in black, for method (B) in dark grey, and for method (C) in light grey. The vertical lines indicate step sizes for which $h\omega$ equals π , 2π , or 3π

implemented in the one-step formulation (2.7)-(2.8) with (2.9). The errors in the x_0 -components are nearly identical for all the methods in the stability range of the Störmer–Verlet method ($h\omega < 2$). Differences between the methods are however visible for larger step sizes. For the other solution components x_1 , \dot{x}_0 , \dot{x}_1 there are pronounced differences in the error behaviour of the methods. All five methods (A)-(E) are considerably more accurate than the Störmer–Verlet method. Figure 2.3 shows the errors of methods (A)-(C) for step sizes beyond the stability range of the Störmer–Verlet method. Methods (A) and (B) lose accuracy when $h\omega$ is near integral multiples of π , a phenomenon that does not occur with method (C).

XIII.2.4 Energy Exchange between Stiff Components

Figure 2.4 shows the energy exchange of the six methods (A)-(F) applied to the Fermi–Pasta–Ulam problem with the same data as in Fig. 2.1. The figures show again the oscillatory energies I_1, I_2, I_3 of the stiff springs, their sum $I = I_1 + I_2 + I_3$ and the total energy $H - 0.8$ as functions of time on the interval $0 \leq t \leq 200$. Only the methods (B), (D) and (F) give a good approximation of the energy exchange between the stiff springs. It will turn out in Sect. XIII.4.2 that a necessary condition for a correct approximation of the energy exchange is $\psi(h\omega)\phi(h\omega) = \text{sinc}(h\omega)$, which is satisfied for method (B). The two-force method (F) satisfies an analogous condition for multi-force methods. The good behaviour of method (D) comes from the fact that here $\psi(h\omega)\phi(h\omega) \approx 0.95 \text{sinc}(h\omega)$ for $h\omega = 1.5$.

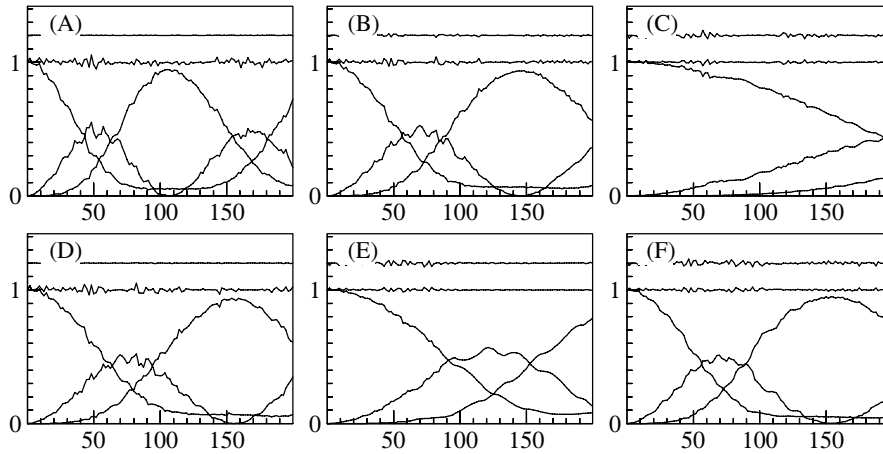


Fig. 2.4. Energy exchange between stiff springs for methods (A)-(F) ($h = 0.03$, $\omega = 50$)

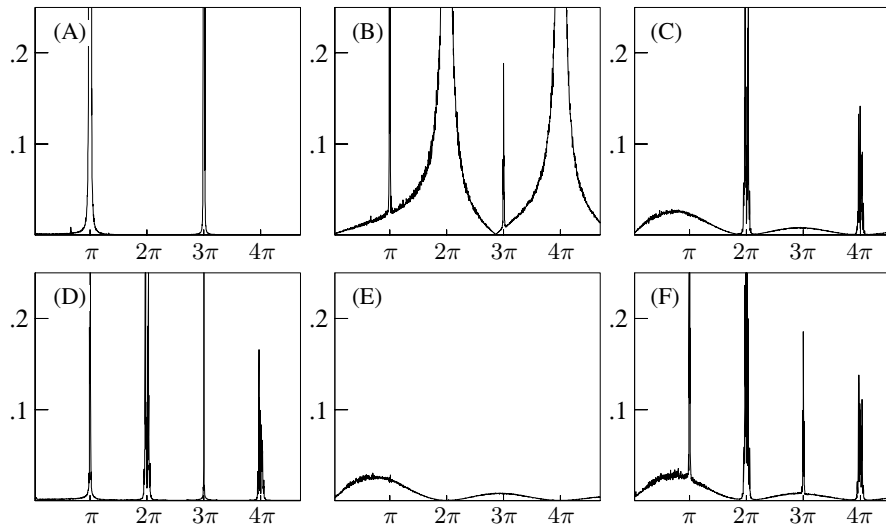


Fig. 2.5. Maximum error of the total energy on the interval $[0, 1000]$ for methods (A) - (F) as a function of $h\omega$ (step size $h = 0.02$)

XIII.2.5 Near-Conservation of Total and Oscillatory Energy

Figure 2.5 shows the maximum error of the total energy H as a function of the scaled frequency $h\omega$ (step size $h = 0.02$). We consider the long time interval $[0, 1000]$. The pictures for the different methods show that in general the total energy is well conserved. Exceptions are near integral multiples of π . Certain methods show a bad energy conservation close to odd multiples of π , other methods close to even multiples of π . Only method (E) shows a uniformly good behaviour for all frequencies. In Fig. 2.6 we show in more detail what happens close to such integral multiples of π .

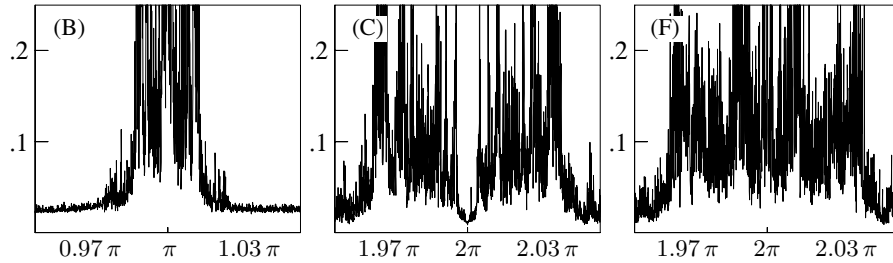


Fig. 2.6. Zoom (close to π or 2π) of the maximum error of the total energy on the interval $[0, 1000]$ for three methods as a function of $h\omega$ (step size $h = 0.02$)

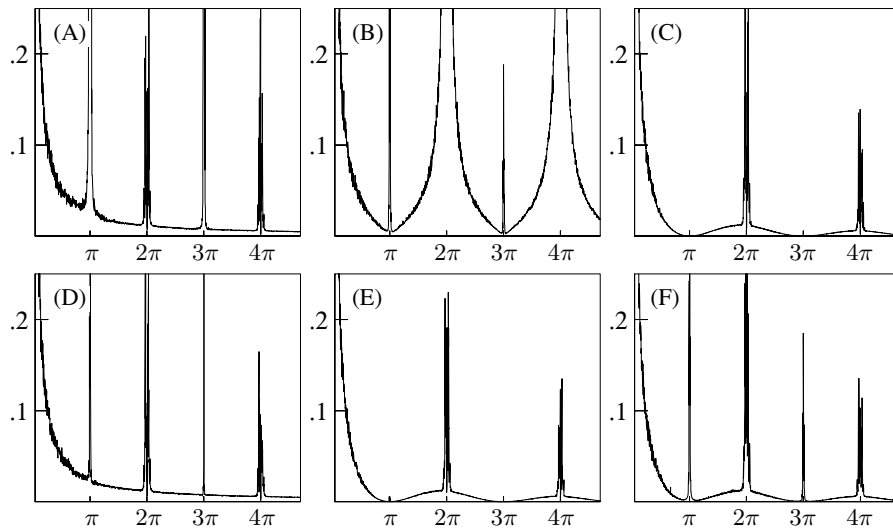


Fig. 2.7. Maximum deviation of the oscillatory energy on the interval $[0, 1000]$ for methods (A) - (F) as a function of $h\omega$ (step size $h = 0.02$)

If there is a difficulty close to π , it is typically in an entire neighbourhood. Close to 2π , the picture is different. Method (C) has good energy conservation for values of $h\omega$ that are very close to 2π , but there are small intervals to the left and to the right, where the error in the total energy is large. Unlike the other methods shown, method (B) has poor energy conservation in rather large intervals around even multiples of π . Methods (A) and (D) conserve the total energy particularly well, for $h\omega$ away from integral multiples of π .

Figure 2.7 shows similar pictures where the total energy H is replaced by the oscillatory energy I (cf. Sect. XIII.2.1). For the exact solution we have $I(t) = \text{Const} + \mathcal{O}(\omega^{-1})$. It is therefore not surprising that this quantity is not well conserved for small values of ω . For larger values of ω , we observe that the methods have difficulties in conserving the oscillatory energy when $h\omega$ is near integral multiples of π . None of the considered methods conserves both quantities H and I uniformly for all values of $h\omega$.

XIII.3 Principal Terms of the Modulated Fourier Expansion

The analytical tool for understanding the above numerical phenomena is provided by *modulated Fourier expansions*, which decompose both the exact and the numerical solution into a slowly varying part and into oscillatory components built up of trigonometric functions multiplied with slowly varying coefficient functions. A comparison of these expansions will serve as a partial substitute for the backward error analysis of Chap. IX, which yields results only for $h\omega \rightarrow 0$ and is not applicable to the situation of $h\omega \geq c > 0$ that is of interest here. In this section we derive the first terms of the modulated Fourier expansion.

XIII.3.1 Decomposition of the Exact Solution

Every solution of the linear equation $\ddot{x} + \Omega^2 x = g(t)$ with Ω of (2.2) can be written as $y(t) + \cos(\omega t) u(t) + \sin(\omega t) v(t) + \mathcal{O}(\omega^{-N})$ (for $\omega \rightarrow \infty$), where $y(t)$, $u(t)$, $v(t)$ are truncated asymptotic expansions in powers of ω^{-1} (see Exercise 4). These functions have the property that all their derivatives are bounded independently of the parameter $\omega \gg 1$. Here and in the following, a *smooth* function is understood to be a function with this property. We may hope to find a similar decomposition for solutions of the nonlinear problem (2.1). So we look for a smooth real-valued function $y(t)$ and a smooth complex-valued function $z(t) = u(t) + iv(t)$ such that the function

$$x_*(t) = y(t) + e^{i\omega t} z(t) + e^{-i\omega t} \bar{z}(t) \quad (3.1)$$

gives a small defect when it is inserted into the differential equation (2.1) and has the given initial values

$$x_*(0) = x(0), \quad \dot{x}_*(0) = \dot{x}(0). \quad (3.2)$$

Under the condition (2.3) the exact solution $x(t)$ has bounded energy, and we may expect the same of the approximation $x_*(t)$, which would then imply $z(t) = \mathcal{O}(\omega^{-1})$. We therefore insert the ansatz (3.1) into the differential equation (2.1) and expand the nonlinearity around the smooth part $y(t)$. With the variables $y = (y_0, y_1)$, $z = (z_0, z_1)$ partitioned according to the blocks of Ω , this gives the expressions

$$\begin{aligned} \ddot{x}_* + \Omega^2 x_* = & \begin{pmatrix} \ddot{y}_0 \\ \ddot{y}_1 + \omega^2 y_1 \end{pmatrix} + e^{i\omega t} \begin{pmatrix} -\omega^2 z_0 + 2i\omega \dot{z}_0 + \ddot{z}_0 \\ 2i\omega \dot{z}_1 + \ddot{z}_1 \end{pmatrix} \\ & + e^{-i\omega t} \begin{pmatrix} -\omega^2 \bar{z}_0 - 2i\omega \dot{\bar{z}}_0 + \ddot{\bar{z}}_0 \\ -2i\omega \dot{\bar{z}}_1 + \ddot{\bar{z}}_1 \end{pmatrix} \end{aligned}$$

and, as long as $z(t) = \mathcal{O}(\omega^{-1})$,

$$\begin{aligned} g(x_*) = & g(y) + g''(y)(z, \bar{z}) + e^{i\omega t} g'(y)z + e^{-i\omega t} g'(y)\bar{z} \\ & + e^{2i\omega t} \frac{1}{2} g''(y)(z, z) + e^{-2i\omega t} \frac{1}{2} g''(y)(\bar{z}, \bar{z}) + \mathcal{O}(\omega^{-3}). \end{aligned}$$

Equations for the Coefficient Functions. We now compare the coefficients of $1, e^{i\omega t}, e^{-i\omega t}$ and require that the dominant terms in these expressions be equal:

$$\begin{aligned}\ddot{y}_0 &= g_0(y) + g_0''(y)(z, \bar{z}) \\ \omega^2 y_1 &= g_1(y) \\ -\omega^2 z_0 &= g_0'(y)z \\ 2i\omega \dot{z}_1 &= g_1'(y)z.\end{aligned}\tag{3.3}$$

This gives a system of differential equations for y_0, z_1 and expresses y_1, z_0 as functions of y_0, z_1 . We note that y_0 evolves on the time scale 1, whereas z_1 changes on the slow time scale ω . As long as $y_0(t)$ stays in a bounded domain and $z_1(t) = \mathcal{O}(\omega^{-1})$, (3.3) implies the bounds

$$y_1(t) = \mathcal{O}(\omega^{-2}), \quad z_0(t) = \mathcal{O}(\omega^{-3}), \quad \dot{z}_1(t) = \mathcal{O}(\omega^{-2}).\tag{3.4}$$

Initial Values. The initial values $y_0(0), \dot{y}_0(0)$ and $z_1(0)$ are obtained from condition (3.2), which gives a system that can be solved by fixed point iteration to yield

$$\begin{aligned}y_0(0) &= x_{0,0} + \mathcal{O}(\omega^{-3}), \quad \dot{y}_0(0) = \dot{x}_{0,0} + \mathcal{O}(\omega^{-2}) \\ 2\operatorname{Re} z_1(0) &= x_{0,1} + \mathcal{O}(\omega^{-2}), \quad -\omega 2\operatorname{Im} z_1(0) = \dot{x}_{0,1} + \mathcal{O}(\omega^{-2}).\end{aligned}\tag{3.5}$$

Defect. As long as $z_1(t) = \mathcal{O}(\omega^{-1})$, the above equations show that the defect

$$d(t) = \ddot{x}_*(t) + \Omega^2 x_*(t) - g(x_*(t))$$

is of the form

$$d(t) = \operatorname{Re} \left(\frac{\omega^{-2} e^{i\omega t} a(t) + \omega^{-2} e^{2i\omega t} b(t) + \mathcal{O}(\omega^{-3})}{\mathcal{O}(\omega^{-2})} \right)\tag{3.6}$$

with smooth functions a, b . Together with (3.3) this also shows that the smooth $\mathcal{O}(\omega^{-2})$ -term $g''(y)(z, \bar{z})$ is the principal term describing the influence of oscillatory solution components on the evolution of smooth components.

Example. To illustrate the approximation of the solution $x(t)$ by $x_*(t)$ of (3.1), we have solved numerically, with high accuracy, the system (3.3) for the FPU problem with the data of Sect. I.5.1. In Figure 3.1 we plot the oscillatory energy $I = I_1 + I_2 + I_3$ with x replaced by the approximation x_* in the definition (2.5) of these quantities. The figure agrees rather well with Figure I.5.2.

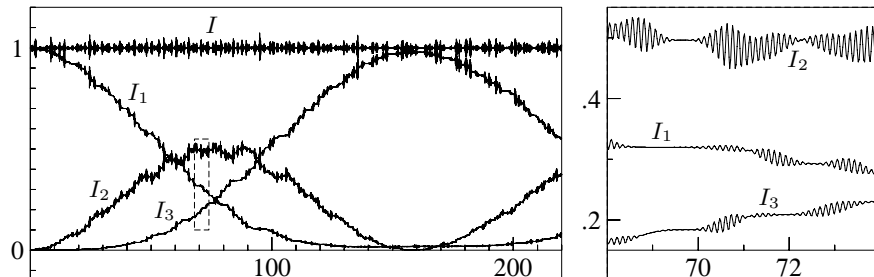


Fig. 3.1. Same experiment as in Fig. I.5.2 for the solution (3.1) of (3.3)

XIII.3.2 Decomposition of the Numerical Solution

For the numerical method (2.6), which solves linear equations $\ddot{x} = -\Omega^2 x$ exactly, we look similarly to the above for a function of the form

$$x_h(t) = y_h(t) + e^{i\omega t} z_h(t) + e^{-i\omega t} \bar{z}_h(t) \quad (3.7)$$

with coefficient functions $y_h(t)$, $z_h(t)$ which are smooth in the sense that all their derivatives are bounded independently of h and ω , such that $x_h(t)$ gives a small defect when inserted into the difference scheme (2.6) and has the correct starting values:

$$x_h(0) = x_0, \quad x_h(h) = x_1. \quad (3.8)$$

Taylor expansion of $z_h(t \pm h)$ at the point t shows, after some calculation,

$$\begin{aligned} \frac{1}{h^2} \left(x_h(t+h) - 2 \cos(h\Omega) x_h(t) + x_h(t-h) \right) &= \left(\sigma_1^2 \omega^2 y_{h,0}(t) + \delta_h^2 y_{h,1}(t) \right) \\ &+ e^{i\omega t} \left(-\sigma_1^2 \omega^2 z_{h,0}(t) + \sigma_2 2i\omega \dot{z}_{h,0}(t) + \cos(h\omega) \ddot{z}_{h,0}(t) + \dots \right) \\ &\quad \sigma_2 2i\omega \dot{z}_{h,1}(t) + \cos(h\omega) \ddot{z}_{h,1}(t) + \dots \end{aligned} \quad (3.9)$$

+ the complex conjugate of the expression in the previous line,

where $y_h(t) = (y_{h,0}(t), y_{h,1}(t))$ and $z_h(t) = (z_{h,0}(t), z_{h,1}(t))$ according to the partitioning in (2.2),

$$\delta_h^2 y_h(t) = \frac{1}{h^2} \left(y_h(t+h) - 2y_h(t) + y_h(t-h) \right)$$

is the symmetric second-order difference quotient, $\sigma_k = \text{sinc}(\frac{1}{2}kh\omega)$, and the dots stand for higher powers of h multiplied by derivatives of $z_{h,0}$ or $z_{h,1}$. Taylor expansion of the nonlinearity now gives

$$\begin{aligned} \Psi g(\Phi x_h) &= \Psi g(\Phi y_h) + \Psi g''(\Phi y_h)(\Phi z_h, \Phi \bar{z}_h) \\ &+ e^{i\omega t} \Psi g'(\Phi y_h) \Phi z_h + e^{-i\omega t} \Psi g'(\Phi y_h) \Phi \bar{z}_h + \dots \end{aligned} \quad (3.10)$$

Modified Equations for the Numerical Coefficient Functions. For the moment we consider the case where the absolute values of σ_1 and σ_2 are bounded from below by a positive constant, so that $h\omega$ is assumed bounded and bounded away from a non-zero integral multiple of π . We also assume $h\omega$ to be bounded away from zero, which is the computational situation of interest. In this case, the first term in each line of each bracket in (3.9) can be considered as the dominant one. We therefore require that the functions y_h, z_h satisfy

$$\begin{aligned} \delta_h^2 y_{h,0} &= g_0(\Phi y_h) + g_0''(\Phi y_h)(\Phi z_h, \Phi \bar{z}_h) \\ \text{sinc}^2(\tfrac{1}{2}h\omega) \omega^2 y_{h,1} &= \psi(h\omega) g_1(\Phi y_h) \\ -\text{sinc}^2(\tfrac{1}{2}h\omega) \omega^2 z_{h,0} &= \frac{\partial g_0}{\partial x_0}(\Phi y_h) z_{h,0} + \frac{\partial g_0}{\partial x_1}(\Phi y_h) \phi(h\omega) z_{h,1} \\ \text{sinc}(h\omega) 2i\omega \dot{z}_{h,1} &= \psi(h\omega) \frac{\partial g_1}{\partial x_0}(\Phi y_h) z_{h,0} + \psi(h\omega) \frac{\partial g_1}{\partial x_1}(\Phi y_h) \phi(h\omega) z_{h,1}. \end{aligned} \quad (3.11)$$

The first equation should be stated more precisely as $y_{h,0}$ being a solution of a modified equation for the Störmer–Verlet method (see Exercise IX.3) applied to the corresponding differential equation:

$$\ddot{y}_{h,0} = \left(1 - \frac{h^2}{12} \frac{d^2}{dt^2}\right) \left(g_0(\Phi y_h) + g_0''(\Phi y_h)(\Phi z_h, \Phi \bar{z}_h)\right),$$

where the time derivatives of $y_{h,1}, z_h$ that result from applying the chain rule are replaced by using the expressions in (3.11). As long as $y_{h,0}(t)$ remains in a bounded domain and $z_{h,1}(t) = \mathcal{O}(\omega^{-1})$, we have again bounds of the same type as for the coefficients of the exact solution:

$$y_{h,1}(t) = \mathcal{O}(\omega^{-2}), \quad z_{h,0}(t) = \mathcal{O}(\omega^{-3}), \quad \dot{z}_{h,1}(t) = \mathcal{O}(\omega^{-2}). \quad (3.12)$$

Initial Values. We next determine the initial values $y_{h,0}(0), \dot{y}_{h,0}(0)$ and $z_{h,1}(0)$ such that $x_h(0)$ and $x_h(h)$ coincide with the starting values $x_0 = x(0)$ and x_1 of the numerical method. We let x_1 be computed from x_0 and \dot{x}_0 via the formula (2.7) with $n = 0$, and we still assume that σ_1 and σ_2 are bounded away from zero. Using (3.11), the condition $x_h(0) = x_0 = (x_{0,0}, x_{0,1})$ then becomes

$$\begin{aligned} x_{0,0} &= y_{h,0}(0) + \mathcal{O}(\omega^{-2} z_{h,1}(0)) \\ x_{0,1} &= z_{h,1}(0) + \bar{z}_{h,1}(0) + \mathcal{O}(\omega^{-2}). \end{aligned} \quad (3.13)$$

The formula for the first component of (2.7), $x_{1,0} - x_{0,0} = h\dot{x}_{0,0} + \frac{1}{2}h^2 g_0(\Phi x_0)$, together with $x_{h,0}(h) - x_{h,0}(0) = h\dot{y}_{h,0}(0) + \frac{1}{2}h^2 g_0(\Phi x_0) + \mathcal{O}(h^3) + \mathcal{O}(\omega^{-2} z_{h,1}(0))$ implies that

$$\dot{x}_{0,0} = \dot{y}_{h,0}(0) + \mathcal{O}(h^2) + \mathcal{O}(\omega^{-1} z_{h,1}(0)). \quad (3.14)$$

For the second component we have from (2.7)

$$x_{1,1} - \cos(h\omega)x_{0,1} = h \operatorname{sinc}(h\omega)\dot{x}_{0,1} + \frac{1}{2}h^2 \psi(h\omega) g_1(\Phi x_0),$$

and by Taylor expansion and (3.11),

$$\begin{aligned} x_{h,1}(h) - \cos(h\omega)x_{h,1}(0) &= (1 - \cos(h\omega))y_{h,1}(0) + \mathcal{O}(h\omega^{-2}) \\ &\quad + i \sin(h\omega)(z_{h,1}(0) - \bar{z}_{h,1}(0)) + \mathcal{O}(h\omega^{-1} z_{h,1}(0)), \end{aligned}$$

where we note the relation $(1 - \cos(h\omega))y_{h,1}(0) = \frac{1}{2}h^2 \psi(h\omega) g_1(\Phi y_h(0))$ by (3.11) and a trigonometric identity. After division by $h \operatorname{sinc} h\omega = \omega^{-1} \sin h\omega$ the above formulas yield

$$\dot{x}_{0,1} = i\omega(z_{h,1}(0) - \bar{z}_{h,1}(0)) + \mathcal{O}(\omega^{-2}) + \mathcal{O}(\omega^{-1} z_{h,1}(0)). \quad (3.15)$$

The four equations (3.13), (3.14), (3.15) constitute a nonlinear system for the four quantities $y_0(0), \dot{y}_0(0), \omega(z_{h,1}(0) + \bar{z}_{h,1}(0))$, and $\omega(z_{h,1}(0) - \bar{z}_{h,1}(0))$. By fixed-point iteration and using the bounded-energy assumption (2.3), we get a locally unique solution for sufficiently small h , with $z_{h,1}(0) = \mathcal{O}(\omega^{-1})$ and hence

$$\begin{aligned}
y_{h,0}(0) &= x_{0,0} + \mathcal{O}(\omega^{-3}), & \dot{y}_{h,0}(0) &= \dot{x}_{0,0} + \mathcal{O}(h^2) \\
2 \operatorname{Re} z_{h,1}(0) &= x_{0,1} + \mathcal{O}(\omega^{-2}), & -\omega 2 \operatorname{Im} z_{h,1}(0) &= \dot{x}_{0,1} + \mathcal{O}(h\omega^{-1}).
\end{aligned} \tag{3.16}$$

Defect. As long as $z_{h,1}(t) = \mathcal{O}(\omega^{-1})$, the defect

$$d_h(t) = \frac{1}{h^2} \left(x_h(t+h) - 2 \cos(h\Omega) x_h(t) + x_h(t-h) \right) - \Psi g(\Phi x_h(t)) \tag{3.17}$$

is of size $\mathcal{O}(h^2)$ by (3.9)–(3.10) and the very construction (3.11) of the coefficient functions. This estimate refers again to the non-resonant case where σ_1, σ_2 are bounded away from zero and hence $h\omega$ is bounded away from non-zero integral multiples of π . The case of $h\omega$ near a multiple of π requires a special treatment and will be considered in the next subsection.

XIII.4 Accuracy and Slow Exchange

A comparison of the principal terms of the modulated Fourier expansions of the numerical and the exact solution gives much insight into the behaviour of the numerical method and the role of the filter functions ψ and ϕ . From this comparison we obtain error bounds over finite time intervals, and we discuss the slow energy exchange between oscillatory components and the slow energy transfer from oscillatory to smooth components which take place on the time scale ω .

XIII.4.1 Convergence Properties on Bounded Time Intervals

As a first application of the modulated Fourier expansion we consider error bounds on bounded time intervals. Second-order convergence estimates for more general equations $\ddot{x} = -Ax + g(x)$ with symmetric positive semi-definite matrix A , uniformly in the (arbitrarily large) eigenfrequencies of A , are given by García-Archilla, Sanz-Serna & Skeel (1999) for the mollified impulse method, by Hochbruck & Lubich (1999a) for Gautschi-type methods, and by Grimm & Hochbruck (2005) for general methods of the class (2.7)–(2.8) with appropriate filter functions. Those results were proved with different techniques. The following bounds on the filter functions ψ and ϕ are needed for second-order error bounds of method (2.6):

$$\begin{aligned}
|\psi(h\omega)| &\leq C_1 \operatorname{sinc}^2(\tfrac{1}{2}h\omega), \\
|\phi(h\omega)| &\leq C_2 |\operatorname{sinc}(\tfrac{1}{2}h\omega)|, \\
|\psi(h\omega)\phi(h\omega)| &\leq C_3 |\operatorname{sinc}(h\omega)|.
\end{aligned} \tag{4.1}$$

Theorem 4.1. *Consider the numerical solution of the system (2.1)–(2.3) by method (2.6) with a step size $h \leq h_0$ (with a sufficiently small h_0 independent of ω) for which $h\omega \geq c_0 > 0$. Let the starting value x_1 be given by (2.7) with $n = 0$. If the conditions (4.1) are satisfied, then the error is bounded by*

$$\|x_n - x(nh)\| \leq C h^2 \quad \text{for } nh \leq T.$$

If only $|\psi(h\omega)| \leq C_0 |\text{sinc}(\frac{1}{2}h\omega)|$ holds instead of (4.1), then the order of convergence reduces to one: $\|x_n - x(nh)\| \leq C h$ for $nh \leq T$. In both cases, C is independent of ω , h and n with $nh \leq T$ and of bounds of solution derivatives, but depends on T , on E of (2.3), on bounds of derivatives of the nonlinearity g , and on C_1, C_2, C_3 or C_0 .

To obtain second-order error bounds uniformly in $h\omega$, condition (4.1) requires a double zero of ψ and a zero of ϕ at even multiples of π , and a zero of ψ or ϕ at odd multiples of π . This is satisfied for the mollified impulse method with $\phi(\xi) = \text{sinc}(\xi)$, for which $\psi(\xi) = \text{sinc}^2(\xi)$. Gautschi-type methods have $\psi(\xi) = \text{sinc}^2(\frac{1}{2}\xi)$, so that the first condition on ψ in (4.1) is trivially satisfied. The conditions on ϕ hold, for example, for $\phi = \text{sinc}$ or for ϕ of (1.20). The original Gautschi method has $\phi = 1$, which does not satisfy the second condition of (4.1), and the Deuffhard/impulse method ($\psi = \text{sinc}$, $\phi = 1$) satisfies only the third condition of (4.1). These latter methods are not of second order uniformly in $h\omega$.

Proof of Theorem 4.1. (a) First we consider the case where $h\omega$ is bounded away from integral multiples of π , so that condition (4.1) is not needed. Comparing the equations (3.3) and (3.11), which determine the modulated Fourier expansion coefficients, shows

$$y_h(t) - y(t) = \mathcal{O}(h^2), \quad z_h(t) - z(t) = \mathcal{O}(h^2)$$

on bounded intervals, and hence

$$x_h(t) - x_*(t) = \mathcal{O}(h^2). \quad (4.2)$$

The variation-of-constants formula (1.8) and a Gronwall-type inequality show that, on bounded intervals, the error $x_*(t) - x(t)$ is of the same magnitude as the defect: by (3.6),

$$x_*(t) - x(t) = \mathcal{O}(\omega^{-2}).$$

The errors $e_n = x_n - x_h(t_n)$ satisfy

$$e_{n+1} - 2 \cos(h\Omega) e_n + e_{n-1} = b_n \quad (4.3)$$

with $b_n = h^2 (\Psi g(\Phi x_n) - \Psi g(\Phi x_h(t_n)) - d_h(t_n))$. This recurrence relation can be solved to yield (Exercise 5)

$$e_{n+1} = -W_{n-1}e_0 + W_n e_1 + \sum_{j=1}^n W_{n-j} b_j \quad (4.4)$$

with

$$W_n = \begin{pmatrix} (n+1)I & 0 \\ 0 & \frac{\sin(n+1)h\omega}{\sin h\omega} I \end{pmatrix}.$$

A discrete Gronwall inequality now yields that on bounded intervals, e_n is of the same magnitude as the defect $d_h(t)$ of (3.17), which is $\mathcal{O}(h^2)$ by the construction of (3.11) and by $z_{h,1} = \mathcal{O}(\omega^{-1})$. Hence,

$$x_n - x_h(t_n) = \mathcal{O}(h^2).$$

Combining these estimates yields the desired second-order error bound.

(b) We now consider the case where $\omega|\text{sinc}(\frac{1}{2}h\omega)| \geq c$ with a sufficiently large constant c , which depends only on bounds of derivatives of g . This condition means that $h\omega$ is outside of an $\mathcal{O}(h)$ neighbourhood of integral multiples of 2π . Under conditions (4.1), the equations (3.11) still give

$$y_{h,1}(t) = \mathcal{O}(\omega^{-2}), \quad z_{h,0}(t) = \mathcal{O}(\omega^{-2}), \quad \dot{z}_{h,1}(t) = \mathcal{O}(\omega^{-2}) \quad (4.5)$$

as long as $z_{h,1}(t) = \mathcal{O}(\omega^{-1})$. Here the first condition of (4.1) gives the bound of $y_{h,1}$, the second one the bound of $z_{h,0}$, and the third one the bound of $\dot{z}_{h,1}$. As in Sect. XIII.3.2, we determine the initial values $y_{h,0}(0)$, $\dot{y}_{h,0}(0)$ and $z_{h,1}(0)$ such that $x_h(0)$ and $x_h(h)$ coincide with the starting values x_0 and x_1 of the numerical method. Using once more (4.1), we obtain a system for the initial values similar to (3.13)–(3.15):

$$\begin{aligned} x_{0,0} &= y_{h,0}(0) + \mathcal{O}(\omega^{-1}z_{h,1}(0)) \\ x_{0,1} &= z_{h,1}(0) + \bar{z}_{h,1}(0) + \mathcal{O}(\omega^{-2}) \\ \dot{x}_{0,0} &= \dot{y}_{h,0}(0) + \mathcal{O}(h) + \mathcal{O}(\omega^{-1}z_{h,1}(0)) \\ \dot{x}_{0,1} &= i\omega(z_{h,1}(0) - \bar{z}_{h,1}(0)) + \mathcal{O}(\omega^{-1}) + \mathcal{O}(z_{h,1}(0)). \end{aligned} \quad (4.6)$$

With the weaker estimates for $z_{h,0}(t)$ and in (4.6) we still obtain estimates for the initial values of the type (3.16) with at most one factor ω^{-1} or h less in the remainder terms. Condition (2.3) implies again $z_1(0) = \mathcal{O}(\omega^{-1})$, which ensures that (4.5) holds for $0 \leq t \leq T$. The defect is then $d_h(t) = \mathcal{O}(h^2)$, and as in part (a) we get the second-order error bound.

(c) Now let $\omega|\text{sinc}(\frac{1}{2}h\omega)| \leq c$, so that $h\omega$ is $\mathcal{O}(h)$ close to a multiple of 2π . In this case we replace the third equation in (3.11) simply by

$$z_{h,0} = 0.$$

Under condition (4.1) we still obtain the bounds (4.5). The initial values are now chosen to satisfy

$$\begin{aligned} x_{0,0} &= y_{h,0}(0) \\ x_{0,1} &= z_{h,1}(0) + \bar{z}_{h,1}(0) + \omega^{-2} \frac{\psi(h\omega)}{\text{sinc}^2(\frac{1}{2}h\omega)} g_1(\Phi x_0) \\ \dot{x}_{0,0} &= \dot{y}_{h,0}(0) \\ \dot{x}_{0,1} &= i\omega(z_{h,1}(0) - \bar{z}_{h,1}(0)). \end{aligned} \quad (4.7)$$

They are then bounded as in (b) and, by the arguments used in the determination of the initial values of Sect. XIII.3.2, yield the estimates $x_h(0) = x_0 + \mathcal{O}(h^3)$

and $x_h(h) = x_1 + \mathcal{O}(h^3)$, and again $z_{h,1}(t) = \mathcal{O}(\omega^{-1})$. Since (4.1) implies $\phi(h\omega)z_{h,1} = \mathcal{O}(\omega^{-2})$ in the present situation of $|\text{sinc}(\frac{1}{2}h\omega)| \leq c\omega^{-1}$, the defect is still $d_h(t) = \mathcal{O}(h^2)$. The bound (4.2) is also seen to hold. Therefore the second-order error bound remains valid in this case.

(d) If only $|\psi(h\omega)| \leq |\text{sinc}(\frac{1}{2}h\omega)|$ holds, then we replace the third equation in (3.11) by $z_{h,0} = 0$. If $\omega|\text{sinc}(\frac{1}{2}h\omega)| \leq 1$, we also set $y_{h,1} = 0$. The defect is then only $d_h(t) = \mathcal{O}(h)$, which yields the first-order error bound. \square

For the velocity approximation, we obtain the following for the method (2.12) or its equivalent formulations.

Theorem 4.2. *Under the conditions of Theorem 4.1, consider the velocity approximation scheme (2.12) with a function ψ_1 satisfying $\psi_1(0) = 1$ and*

$$|\psi_1(h\omega)| \leq C'_1 |\text{sinc}(\frac{1}{2}h\omega)|. \quad (4.8)$$

Let the starting values satisfy $\dot{x}_0 = \dot{x}(0)$ and $\dot{x}_1 = \dot{x}(h) + h \sin(h\Omega)a_1 + \mathcal{O}(h^2)$ with $a_1 = \mathcal{O}(1)$. Then, the error in the velocities is bounded by

$$\|\dot{x}_n - \dot{x}(nh)\| \leq C h \quad \text{for } nh \leq T,$$

where C is independent of ω , h and n with $nh \leq T$ and of bounds of solution derivatives, but depends on T , on E of (2.3), on bounds of derivatives of the nonlinearity g , and on C_1, C_2, C_3 and C'_1 .

Proof. (a) By the variation-of-constants formula (1.8), the exact solution satisfies

$$\begin{aligned} & \dot{x}(t+h) - 2\cos(h\Omega)\dot{x}(t) + \dot{x}(t-h) \\ &= \int_0^h \cos((h-s)\Omega) \left(g(x(t+s)) - g(x(t-s)) \right) ds. \end{aligned}$$

With the modulated Fourier expansion, we write the exact solution as

$$x(t) = y(t) + e^{i\omega t}z(t) + e^{-i\omega t}\bar{z}(t) + \mathcal{O}(\omega^{-2})$$

to obtain

$$\begin{aligned} & g(x(t+s)) - g(x(t-s)) \\ &= g'(y(t)) \left(2s\dot{y}(t) - 4\sin(\omega s) \text{Im}(e^{i\omega t}z(t)) + \mathcal{O}(s^2) + \mathcal{O}(\omega^{-2}) \right). \end{aligned}$$

Using the bounds (3.4), abbreviating $g_{i,j} = \partial g_i / \partial x_j$ and omitting the arguments t and $y(t)$ on the right-hand side, we therefore have

$$\begin{aligned} & \dot{x}(t+h) - 2\cos(h\Omega)\dot{x}(t) + \dot{x}(t-h) \\ &= \left(\begin{aligned} & h^2 g_{0,0} \dot{y}_0 - 2h^2 \text{sinc}^2(\frac{1}{2}h\omega) \omega g_{0,1} \text{Im}(e^{i\omega t}z_1) + \mathcal{O}(h^3) \\ & h^2 \text{sinc}^2(\frac{1}{2}h\omega) g_{1,0} \dot{y}_0 - 2h^2 \text{sinc}(h\omega) \omega g_{1,1} \text{Im}(e^{i\omega t}z_1) + \mathcal{O}(h^3) \end{aligned} \right). \end{aligned}$$

We now use the discrete variation-of-constants formula (4.4) and partial summation. For example, the expression

$$\sum_{j=1}^n \frac{\sin(n+1-j)h\omega}{\sin h\omega} \frac{1}{2} h^2 \text{sinc}^2(\frac{1}{2}h\omega) g_{1,0}(y(jh)) \dot{y}_0(jh)$$

is seen to be $\mathcal{O}(h)$ uniformly in $h\omega$ and for $nh \leq T$ by partial summation, using that the function $g_{1,0}(y(t))\dot{y}_0(t)$ has a bounded derivative and that

$$\frac{\sin(\frac{1}{2}h\omega)}{\sin(h\omega)} \sum_{j=1}^k \sin(jh\omega) = \mathcal{O}(k) .$$

In this way we obtain

$$\begin{aligned} \dot{x}(nh) = & -W_{n-1} \dot{x}(0) + W_n \dot{x}(h) \\ & + \left(h \sum_{j=1}^n (n+1-j) \frac{h}{0} g_{0,0}(y(jh)) \dot{y}_0(jh) \right) + \mathcal{O}(h) . \end{aligned} \quad (4.9)$$

(b) For the numerical approximation we proceed similarly. Inserting the modulated Fourier expansion of the numerical solution,

$$x_n = y_h(t) + e^{i\omega t} z_h(t) + e^{-i\omega t} \bar{z}_h(t) + \mathcal{O}(h^2) \quad \text{for } t = nh \leq T ,$$

into the numerical scheme, we have with (3.12) or (4.5)

$$\begin{aligned} & \dot{x}_{n+1} - 2 \cos(h\omega) \dot{x}_n + \dot{x}_{n-1} \\ & = h^2 \left(\begin{aligned} & g_{0,0} \dot{y}_{h,0} - 2 \phi(h\omega) \text{sinc}(h\omega) \omega g_{0,1} \text{Im}(e^{i\omega t} z_{h,1}) + \mathcal{O}(h) \\ & \psi_1(h\omega) g_{1,0} \dot{y}_{h,0} - 2 (\psi_1 \phi)(h\omega) \text{sinc}(h\omega) \omega g_{1,1} \text{Im}(e^{i\omega t} z_{h,1}) + \mathcal{O}(h) \end{aligned} \right) \end{aligned}$$

where the functions $g_{i,j}$ are evaluated at $\Phi y_h(t)$ and the argument $t = nh$ is to be inserted in $\dot{y}_{h,0}$ and $z_{h,1}$. Under the condition (4.8) on ψ_1 , we obtain as in (4.9)

$$\begin{aligned} \dot{x}_n = & -W_{n-1} \dot{x}_0 + W_n \dot{x}_1 \\ & + \left(h \sum_{j=1}^n (n+1-j) \frac{h}{0} g_{0,0}(\Phi y_h(jh)) \dot{y}_{h,0}(jh) \right) + \mathcal{O}(h) . \end{aligned} \quad (4.10)$$

Since we know from the estimates (3.12) and from the proof of Theorem 4.1 that $\Phi y_h(t) = y(t) + \mathcal{O}(h^2)$ and $\dot{y}_h(t) = \dot{y}(t) + \mathcal{O}(h^2)$, a comparison of (4.9) and (4.10) gives the result. \square

XIII.4.2 Intra-Oscillatory and Oscillatory-Smooth Exchanges

In this subsection we turn to the approximation of slow effects that take place on the time scale ω . Since solutions may depart from each other exponentially, we

cannot expect to obtain small point-wise error bounds on such a time scale. Instead, we take recourse to a kind of formal backward error analysis where we require that the equations determining the modulated Fourier expansion coefficients for the numerical method be small perturbations of those for the exact solution. It may be expected that methods with this property – *ceteribus paris* – show a better long-time behaviour, and this is indeed confirmed by the numerical experiments.

In the Fermi–Pasta–Ulam model, the oscillatory energy of the j th stiff spring is

$$I_j = \frac{1}{2} \dot{x}_{1,j}^2 + \frac{1}{2} \omega^2 x_{1,j}^2 ,$$

where $x_{1,j}$ is the j th component of the lower block x_1 of x . In terms of the modulated Fourier expansion, this is approximately, up to $\mathcal{O}(\omega^{-1})$,

$$I_j \approx \frac{1}{2} |i\omega z_{1,j} e^{i\omega t} - i\omega \bar{z}_{1,j} e^{-i\omega t}|^2 + \frac{1}{2} \omega^2 |z_{1,j} e^{i\omega t} + \bar{z}_{1,j} e^{-i\omega t}|^2 = 2\omega^2 |z_{1,j}|^2 .$$

The energy exchange between stiff springs as shown in Fig. 2.1 is thus caused by the slow evolution of z_1 determined by (3.3). This should be modeled correctly by the numerical method.

The term $g_0''(y)(z, \bar{z})$ in the differential equation for y_0 in (3.3) is the dominant term by which the oscillations of the stiff springs exert an influence on the smooth motion. A correct incorporation of this term in the numerical method is desirable.

Upon eliminating y_1 and z_0 in (3.3), the differential equations for y_0 and z_1 become, up to $\mathcal{O}(\omega^{-3})$ perturbations on the right-hand sides,

$$\begin{aligned} \ddot{y}_0 &= g_0(y_0, \omega^{-2} g_1(y_0, 0)) + \frac{\partial^2 g_0}{\partial x_1^2}(y_0, 0)(z_1, \bar{z}_1) \\ 2i\omega \dot{z}_1 &= \frac{\partial g_1}{\partial x_1}(y_0, 0) z_1 . \end{aligned} \quad (4.11)$$

This is to be compared with the analogous equations for the modulated Fourier expansion of the numerical method, which follow from (3.11):

$$\begin{aligned} \delta_h^2 y_{h,0} &= g_0(y_{h,0}, \gamma \omega^{-2} g_1(y_{h,0}, 0)) + \beta \frac{\partial^2 g_0}{\partial x_1^2}(y_{h,0}, 0)(z_{h,1}, \bar{z}_{h,1}) \\ 2i\omega \dot{z}_{h,1} &= \alpha \frac{\partial g_1}{\partial x_1}(y_{h,0}, 0) z_{h,1} \end{aligned} \quad (4.12)$$

with

$$\alpha = \frac{(\psi\phi)(h\omega)}{\text{sinc}(h\omega)}, \quad \beta = \phi(h\omega)^2, \quad \gamma = \frac{(\psi\phi)(h\omega)}{\text{sinc}^2(\frac{1}{2}h\omega)}. \quad (4.13)$$

The differential equation for $z_{h,1}$ is consistent with that for z_1 only if $\alpha = 1$, i.e.,

$$\psi(h\omega) \phi(h\omega) = \text{sinc}(h\omega) . \quad (4.14)$$

Among all the methods (2.6) considered, only the Deuffhard/impulse method ($\psi = \text{sinc}$, $\phi = 1$) satisfies this condition. For this method we indeed observe a qualitatively correct approximation of the energy exchange between stiff springs in

Fig. 2.4, but we have also seen that the energy conservation of this method is very sensitive to near-resonances.

A correct modeling of the slow oscillatory–smooth transfer would in addition require $\beta = 1$ and possibly $\gamma = 1$. For general $h\omega$ the condition $\gamma = 1$ is, however, incompatible with (4.14).

Multi-force methods (2.13) offer a way out of these difficulties. For such methods, the coefficients of the modulated Fourier expansion satisfy (4.12) with (4.13) replaced by

$$\begin{aligned}\alpha &= \frac{\sum_j \psi_j(h\omega) \phi_j(h\omega)}{\text{sinc}(h\omega)}, \quad \beta = \sum_j \psi_j(0) \phi_j(h\omega)^2, \\ \gamma &= \sum_j \psi_j(0) \phi_j(h\omega) \frac{\sum_k \psi_k(h\omega)}{\text{sinc}^2(\frac{1}{2}h\omega)}.\end{aligned}\quad (4.15)$$

The two-force method (1.23) with (1.25) has $\alpha = \beta = \gamma = 1$ as desired.

XIII.5 Modulated Fourier Expansions

The decomposition of the exact and the numerical solution into modulated exponentials and a remainder, as derived in Sect. XIII.3, was found useful for understanding several important aspects of the numerical behaviour. Those few terms are, however, not sufficient for explaining the long-time near-conservation of the total and the oscillatory energy. The expansion can be made more accurate by adding further terms $e^{\pm 2i\omega t}$, $e^{\pm 3i\omega t}$ etc. multiplied by slowly varying functions. This leads to an asymptotic expansion which we call the *modulated Fourier expansion*. This expansion is constructed in the present section, following Hairer & Lubich (2000a). (In that paper the modulated Fourier expansion was called the frequency expansion.)

XIII.5.1 Expansion of the Exact Solution

The following theorem extends the construction of Sect. XIII.3.1 to arbitrary order in ω^{-1} .

Theorem 5.1. *Consider a solution $x(t)$ of (2.1) which satisfies the bounded-energy condition (2.3) and stays in a compact set K for $0 \leq t \leq T$. Then, the solution admits an expansion*

$$x(t) = y(t) + \sum_{0 < |k| < N} e^{ik\omega t} z^k(t) + R_N(t) \quad (5.1)$$

for arbitrary $N \geq 2$, where the remainder term and its derivative are bounded by

$$R_N(t) = \mathcal{O}(\omega^{-N-2}) \quad \text{and} \quad \dot{R}_N(t) = \mathcal{O}(\omega^{-N-1}) \quad \text{for} \quad 0 \leq t \leq T. \quad (5.2)$$

The real-valued functions $y = (y_0, y_1)$ and the complex-valued functions $z^k = (z_0^k, z_1^k)$ together with all their derivatives (up to arbitrary order M) are bounded by

$$\begin{aligned} y_0 &= \mathcal{O}(1), & z_0^1 &= \mathcal{O}(\omega^{-3}), & z^k &= \mathcal{O}(\omega^{-k-2}) \\ y_1 &= \mathcal{O}(\omega^{-2}), & z_1^1 &= \mathcal{O}(\omega^{-1}), \end{aligned} \quad (5.3)$$

for $k = 2, \dots, N-1$. Moreover, $z^{-k} = \overline{z^k}$ for all k . These functions are unique up to terms of size $\mathcal{O}(\omega^{-N-2})$. The constants symbolized by the \mathcal{O} -notation are independent of ω and t with $0 \leq t \leq T$ (but they depend on N , T , on E of (2.3), on bounds of the derivatives of the nonlinearity $g(x)$ on K , and on the maximum order M of considered derivatives).

Proof. We set

$$x_*(t) = y(t) + \sum_{0 < |k| < N} e^{ik\omega t} z^k(t) \quad (5.4)$$

and determine the smooth functions $y(t)$, $z(t) = z^1(t)$, and $z^2(t), \dots, z^{N-1}(t)$ such that $x_*(t)$ inserted into the differential equation (2.1) has a small defect, of size $\mathcal{O}(\omega^{-N})$. To this end we expand $g(x_*(t))$ around $y(t)$ and compare the coefficients of $e^{ik\omega t}$. With the notation $g^{(m)}(y)z^\alpha = g^{(m)}(y)(z^{\alpha_1}, \dots, z^{\alpha_m})$ for a multi-index $\alpha = (\alpha_1, \dots, \alpha_m)$, there results the following system of differential equations:

$$\begin{pmatrix} \ddot{y}_0 \\ \omega^2 y_1 \end{pmatrix} + \begin{pmatrix} 0 \\ \ddot{y}_1 \end{pmatrix} = g(y) + \sum_{s(\alpha)=0} \frac{1}{m!} g^{(m)}(y) z^\alpha \quad (5.5)$$

$$\begin{pmatrix} -\omega^2 z_0 \\ 2i\omega \dot{z}_1 \end{pmatrix} + \begin{pmatrix} 2i\omega \dot{z}_0 + \ddot{z}_0 \\ \ddot{z}_1 \end{pmatrix} = \sum_{s(\alpha)=1} \frac{1}{m!} g^{(m)}(y) z^\alpha \quad (5.6)$$

$$\begin{pmatrix} -k^2 \omega^2 z_0^k \\ (1 - k^2) \omega^2 z_1^k \end{pmatrix} + \begin{pmatrix} 2ki\omega \dot{z}_0^k + \ddot{z}_0^k \\ 2ki\omega \dot{z}_1^k + \ddot{z}_1^k \end{pmatrix} = \sum_{s(\alpha)=k} \frac{1}{m!} g^{(m)}(y) z^\alpha. \quad (5.7)$$

Here the sums range over all $m \geq 1$ and all multi-indices $\alpha = (\alpha_1, \dots, \alpha_m)$ with integers α_j satisfying $0 < |\alpha_j| < N$, which have a given sum $s(\alpha) = \sum_{j=1}^m \alpha_j$.

For large ω , the dominating terms in these differential equations are given by the left-most expressions. However, since the central terms involve higher derivatives, we are confronted with singular perturbation problems. We are interested in smooth functions y, z, z^k that satisfy the system up to a defect of size $\mathcal{O}(\omega^{-N})$. In the spirit of Euler's derivation of the Euler-Maclaurin summation formula (see e.g. Hairer & Wanner 1997) we remove the disturbing higher derivatives by using iteratively the differentiated equations (5.5)-(5.7). This leads to a system

$$\begin{aligned} \ddot{y}_0 &= \mathcal{F}_0(\dot{y}_0, y, z^1, \dots, z^{N-1}, \omega^{-1}), & \dot{z}_1 &= \omega^{-1} \mathcal{F}_1(\dot{y}_0, y, z^1, \dots, z^{N-1}, \omega^{-1}) \\ z_0 &= \omega^{-2} \mathcal{G}_0(\dot{y}_0, y, z^1, \dots, z^{N-1}, \omega^{-1}), & y_1 &= \omega^{-2} \mathcal{G}_1(\dot{y}_0, y, z^1, \dots, z^{N-1}, \omega^{-1}) \\ z_0^k &= \omega^{-2} \mathcal{G}_0^k(\dot{y}_0, y, z^1, \dots, z^{N-1}, \omega^{-1}), & z_1^k &= \omega^{-2} \mathcal{G}_1^k(\dot{y}_0, y, z^1, \dots, z^{N-1}, \omega^{-1}) \end{aligned}$$

where $\mathcal{F}_j, \mathcal{G}_j, \mathcal{G}_j^k$ are formal series in powers of ω^{-1} . Since we get formal algebraic relations for y_1, z_0, z^k , we can further eliminate these variables in the functions $\mathcal{F}_j, \mathcal{G}_j, \mathcal{G}_j^k$. We finally obtain for y_1, z_1, z^k the algebraic relations

$$\begin{aligned} z_0 &= \omega^{-2}(G_{00}(y_0, \dot{y}_0, z_1) + \omega^{-1}G_{01}(y_0, \dot{y}_0, z_1) + \dots) \\ y_1 &= \omega^{-2}(G_{10}(y_0, \dot{y}_0, z_1) + \omega^{-1}G_{11}(y_0, \dot{y}_0, z_1) + \dots) \\ z_0^k &= \omega^{-2}(G_{00}^k(y_0, \dot{y}_0, z_1) + \omega^{-1}G_{01}^k(y_0, \dot{y}_0, z_1) + \dots) \\ z_1^k &= \omega^{-2}(G_{10}^k(y_0, \dot{y}_0, z_1) + \omega^{-1}G_{11}^k(y_0, \dot{y}_0, z_1) + \dots) \end{aligned} \quad (5.8)$$

and a system of real second-order differential equations for y_0 and complex first-order differential equations for z_1 :

$$\begin{aligned} \ddot{y}_0 &= F_{00}(y_0, \dot{y}_0, z_1) + \omega^{-1}F_{01}(y_0, \dot{y}_0, z_1) + \dots \\ \dot{z}_1 &= \omega^{-1}(F_{10}(y_0, \dot{y}_0, z_1) + \omega^{-1}F_{11}(y_0, \dot{y}_0, z_1) + \dots). \end{aligned} \quad (5.9)$$

At this point we can forget the above derivation and take it as a motivation for the ansatz (5.8)-(5.9), which is truncated after the $\mathcal{O}(\omega^{-N})$ terms. We insert this ansatz and its first and second derivatives into (5.5)-(5.7) and compare powers of ω^{-1} . This yields recurrence relations for the functions F_{jl}^k, G_{jl}^k , which in addition show that these functions together with their derivatives are all bounded on compact sets.

We determine initial values for (5.9) such that the function $x_*(t)$ of (5.4) satisfies $x_*(0) = x_0$ and $\dot{x}_*(0) = \dot{x}_0$. Because of the special ansatz (5.8)-(5.9), this gives a system which, by fixed-point iteration, yields (locally) unique initial values $y_0(0), \dot{y}_0(0), z_1(0)$ satisfying (3.5). The assumption (2.3) implies that $z_1(0) = \mathcal{O}(\omega^{-1})$. It further follows from the boundedness of F_{1l} that $z_1(t) = \mathcal{O}(\omega^{-1})$ for $0 \leq t \leq T$. Going back to (5.7), it is seen that the functions G_{jl}^k contain at least k times the factor z_1 . This implies the stated bounds for all other functions.

It remains to estimate the error $R_N(t) = x(t) - x_*(t)$. For this we consider the solution of (5.8)-(5.9) with the above initial values. By construction, these functions satisfy the system (5.5)-(5.7) up to a defect of $\mathcal{O}(\omega^{-N})$. This gives a defect of size $\mathcal{O}(\omega^{-N})$ when the function $x_*(t)$ of (5.4) is inserted into (2.1). On a finite time interval $0 \leq t \leq T$, this implies $R_N(t) = \mathcal{O}(\omega^{-N})$ and $\dot{R}_N(t) = \mathcal{O}(\omega^{-N})$. To obtain the slightly sharper bounds (5.2), we apply the above proof with N replaced by $N + 2$ and use the bounds (5.3) for z^N and z^{N+1} . \square

XIII.5.2 Expansion of the Numerical Solution

Does the numerical solution of (2.1) have a modulated Fourier expansion similar to the analytical solution? This may of course be expected, but in Sect. XIII.3.2 we encountered difficulties in constructing the first terms of the expansion in the situation of a numerical resonance where $h\omega$ is close to an integral multiple of π . We therefore confine the discussion to the non-resonant case. We assume that h and ω^{-1} lie in a subregion of the (h, ω^{-1}) -plane of small parameters for which there exists a positive constant c such that

$$|\sin(\frac{1}{2}kh\omega)| \geq c\sqrt{h} \quad \text{for } k = 1, \dots, N, \text{ with } N \geq 2. \quad (5.10)$$

This condition implies that $h\omega$ is outside an $\mathcal{O}(\sqrt{h})$ neighbourhood of integral multiples of π . For given h and ω , the condition imposes a restriction on N . In the following, N is a fixed integer such that (5.10) holds. There is the following numerical analogue of Theorem 5.1.

Theorem 5.2. *Consider the numerical solution of the system (2.1)–(2.3) by method (2.6) with step size h . Let the starting value x_1 be given by (2.7) with $n = 0$. Assume $h\omega \geq c_0 > 0$, the non-resonance condition (5.10), and the bounds (4.1) for $\psi(h\omega)$ and $\phi(h\omega)$. Then, the numerical solution admits an expansion*

$$x_n = y_h(t) + \sum_{0 < |k| < N} e^{ik\omega t} z_h^k(t) + R_{h,N}(t) \quad (5.11)$$

uniformly for $0 \leq t = nh \leq T$. The remainder term is of the form

$$R_{h,N}(t) = t^2 h^N \Psi r(t) \quad \text{with } r(t) = \mathcal{O}(\phi(h\omega)^N + h^m), \quad (5.12)$$

where $m \geq 0$ can be chosen arbitrarily. The coefficient functions together with all their derivatives (up to some arbitrarily fixed order) are bounded by

$$\begin{aligned} y_{h,0} &= \mathcal{O}(1), & z_{h,0}^1 &= \mathcal{O}(\omega^{-2}), & z_{h,0}^k &= \mathcal{O}(\omega^{-k}), \\ y_{h,1} &= \mathcal{O}(\omega^{-2}), & z_{h,1}^1 &= \mathcal{O}(\omega^{-1}), & z_{h,1}^k &= \mathcal{O}(\omega^{-k}) \end{aligned} \quad (5.13)$$

for $k = 2, \dots, N-1$. Moreover, $z_h^{-k} = \overline{z_h^k}$ for all k . The constants symbolized by the \mathcal{O} -notation are independent of ω and h with (5.10), but they depend on E , N , m , c , and T .

The proof covers the remainder of this subsection. It constructs a function

$$x_h(t) = y_h(t) + \sum_{0 < |k| < N} e^{ik\omega t} z_h^k(t) \quad (5.14)$$

with smooth coefficient functions $y_h(t)$ and $z_h^k(t)$, which has a small defect when it is inserted into the numerical scheme (2.6). The following functional calculus is convenient for determining the coefficient functions.

Functional Calculus. Let f be an entire complex function bounded by $|f(\zeta)| \leq C e^{\gamma|\zeta|}$. Then,

$$f(hD)x(t) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} h^k x^{(k)}(t)$$

converges for every function x which is analytic in a disk of radius $r > \gamma h$ around t . If f_1 and f_2 are two such entire functions, then

$$f_1(hD)f_2(hD)x(t) = (f_1 f_2)(hD)x(t)$$

whenever both sides exist. We note $(hD)^k x(t) = h^k x^{(k)}(t)$ for $k = 0, 1, 2, \dots$ and $\exp(hD)x(t) = x(t+h)$.

We next consider the application of such an operator to functions of the form $e^{i\omega t} z(t)$. By Leibniz' rule of calculus we have $(hD)^k e^{i\omega t} z(t) = e^{i\omega t} (hD + ih\omega)^k z(t)$. After a short calculation this yields

$$f(hD)e^{i\omega t} z(t) = e^{i\omega t} f(hD + ih\omega)z(t) \quad (5.15)$$

where $f(hD + ih\omega)z(t) = \sum_{k=0}^{\infty} f^{(k)}(ih\omega)/k! \cdot h^k z^{(k)}(t)$.

An N -times continuously differentiable function x is replaced by its Taylor polynomial of degree $N-1$ at t , and $f(hD)x(t)$ is then considered up to $\mathcal{O}(h^N)$.

Modified Equations for the Coefficient Functions. The difference operator of the numerical method becomes in this notation

$$x(t+h) - 2\cos h\Omega x(t) + x(t-h) = (e^{hD} - 2\cos h\Omega + e^{-hD})x(t).$$

We factorize this operator as

$$\begin{aligned} \mathcal{L}(hD) &:= e^{hD} - 2\cos h\Omega + e^{-hD} = 2(\cos(ihD) - \cos h\Omega) \\ &= 4 \sin\left(\frac{1}{2}h\Omega + \frac{1}{2}ihD\right) \sin\left(\frac{1}{2}h\Omega - \frac{1}{2}ihD\right). \end{aligned} \quad (5.16)$$

The function $x_h(t)$ of (5.14) should formally (up to $\mathcal{O}(h^{N+2})$) satisfy the difference scheme

$$\mathcal{L}(hD)x_h(t) = h^2\Psi g(\Phi x_h(t)). \quad (5.17)$$

We insert the ansatz (5.14), expand the right-hand side into a Taylor series around $\Phi y_h(t)$, and compare the coefficients of $e^{ik\omega t}$. This yields the following formal equations for the functions $y_h(t)$ and $z_h^k(t)$:

$$\begin{aligned} \mathcal{L}(hD)y_h &= h^2\Psi \left(g(\Phi y_h) + \sum_{s(\alpha)=0} \frac{1}{m!} g^{(m)}(\Phi y_h)(\Phi z_h)^\alpha \right) \\ \mathcal{L}(hD + ikh\omega)z_h^k &= h^2\Psi \sum_{s(\alpha)=k} \frac{1}{m!} g^{(m)}(\Phi y_h)(\Phi z_h)^\alpha. \end{aligned} \quad (5.18)$$

Here, $\alpha = (\alpha_1, \dots, \alpha_m)$ is a multi-index as in the proof of Theorem 5.1, $s(\alpha) = \sum_{j=1}^m \alpha_j$, and $(\Phi z)^\alpha$ is an abbreviation for the m -tuple $(\Phi z^{\alpha_1}, \dots, \Phi z^{\alpha_m})$. To get smooth functions $y_h(t)$ and $z_h^k(t)$ which solve (5.18) up to a small defect, we look at the dominating terms in the Taylor expansions of $\mathcal{L}(hD)$ and $\mathcal{L}(hD + ikh\omega)$. With the abbreviations $s_k = \sin(\frac{1}{2}kh\omega)$ and $c_k = \cos(\frac{1}{2}kh\omega)$ we obtain

$$\begin{aligned} \mathcal{L}(hD) &= \begin{pmatrix} 0 & 0 \\ 0 & 4s_1^2 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (ihD)^2 + \dots \\ \mathcal{L}(hD + ih\omega) &= \begin{pmatrix} -4s_1^2 & 0 \\ 0 & 0 \end{pmatrix} + 2s_2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (ihD) \end{aligned}$$

$$\begin{aligned}
& -c_2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (ihD)^2 + \dots \quad (5.19) \\
\mathcal{L}(hD + ikh\omega) = & \begin{pmatrix} -4s_k^2 & 0 \\ 0 & -4s_{k-1}s_{k+1} \end{pmatrix} + 2s_{2k} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (ihD) \\
& -c_{2k} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} (ihD)^2 + \dots
\end{aligned}$$

Construction of the Coefficient Functions. Under the non-resonance condition (5.10), the first non-vanishing coefficients in (5.19) are the dominant ones, and the derivation of the defining relations for y_h and z_h^k is the same as for the analytical solution in Theorem 5.1; see also part (b) of the proof of Theorem 4.1. We insert (5.19) into (5.18) and we eliminate recursively the higher derivatives. This motivates the following ansatz for the computation of the functions y_h and z_h^k :

$$\begin{aligned}
\ddot{y}_{h,0} &= f_{00}(\cdot) + \sqrt{h} f_{01}(\cdot) + h f_{02}(\cdot) + \dots \\
\dot{z}_{h,1}^1 &= \frac{\psi(h\omega)h}{s_2} \left(f_{10}(\cdot) + \sqrt{h} f_{11}(\cdot) + \dots \right) \\
z_{h,0}^1 &= \frac{h^2}{s_1^2} \left(g_{00}^1(\cdot) + \sqrt{h} g_{01}^1(\cdot) + \dots \right) \\
y_{h,1} &= \frac{\psi(h\omega)h^2}{s_1^2} \left(g_{10}(\cdot) + \sqrt{h} g_{11}(\cdot) + \dots \right) \quad (5.20) \\
z_{h,0}^k &= \frac{h^2}{s_k^2} \left(g_{00}^k(\cdot) + \sqrt{h} g_{01}^k(\cdot) + \dots \right) \\
z_{h,1}^k &= \frac{\psi(h\omega)h^2}{s_{k+1}s_{k-1}} \left(g_{10}^k(\cdot) + \sqrt{h} g_{11}^k(\cdot) + \dots \right),
\end{aligned}$$

for $k = 2, \dots, N-1$, where the functions depend smoothly on the variables $y_{h,0}, \dot{y}_{h,0}, \phi(h\omega)z_{h,1}^1$ and on the bounded parameters $\sqrt{h}/s_k, s_k, c_k, \psi(h\omega)$ and $(h\omega)^{-1}$. Inserting this ansatz and its derivatives into (5.18) and comparing powers of \sqrt{h} yields recurrence relations for the functions f_{jl}^k, g_{jl}^k . The functions g_{jl}^k (for $k \geq 1$) contain at least k times the factor $\phi(h\omega)z_{h,1}^1$, and f_{1l} contains this factor at least once. Since the series in (5.20) need not converge, we truncate them after the h^{N+m+2} terms.

Initial Values. The conditions $x_h(0) = x_0$ and $x_h(h) = x_1$ determine the initial values $y_{h,0}(0), \dot{y}_{h,0}(0)$ and $z_{h,1}(0)$ in the same way as in Sect. XIII.3.2. Condition (4.1) yields again (4.6), and (2.3) then implies $z_{h,1}(0) = \mathcal{O}(\omega^{-1})$.

Defect. It follows from (4.1) that $h\psi(h\omega)\phi(h\omega)/s_2 = \mathcal{O}(\omega^{-1})$, so that $\dot{z}_{h,1}^1 = \mathcal{O}(\omega^{-1}z_{h,1}^1)$ by (5.20). This implies $z_{h,1}^1(t) = \mathcal{O}(\omega^{-1})$ for $t \leq T$. The other estimates (5.13) are directly obtained from (5.20), which indeed yields the following more refined bounds for the coefficient functions together with their derivatives:

$$\begin{aligned}
y_{h,0} &= \mathcal{O}(1), & y_{h,1} &= \mathcal{O}(\omega^{-2}) \\
z_{h,0}^1 &= \mathcal{O}(\omega^{-3}/\sqrt{h}), & z_{h,1}^1 &= \mathcal{O}(\omega^{-1}), & \dot{z}_{h,1}^1 &= \mathcal{O}(\omega^{-2}) \\
z_{h,0}^k &= \mathcal{O}(h\phi(h\omega)^k\omega^{-k}), & z_{h,1}^k &= \mathcal{O}(h\psi(h\omega)\phi(h\omega)^k\omega^{-k}).
\end{aligned} \tag{5.21}$$

Consequently, the values $x_h(nh)$ inserted into the numerical scheme (2.6) yield a defect of size $\mathcal{O}(h^{N+2})$:

$$\begin{aligned}
x_h(t+h) - 2\cos(h\Omega)x_h(t) + x_h(t-h) &= \\
&= h^2\Psi\left(g(\Phi x_h(t)) + \mathcal{O}(\phi(h\omega)^N\omega^{-N} + h^{N+m})\right).
\end{aligned} \tag{5.22}$$

Standard convergence estimates then show that, on bounded time intervals, $x_n - x_h(nh)$ is of size $\mathcal{O}(t^2h^N)$ and actually satisfies the finer estimate (5.12). This completes the proof of Theorem 5.2. \square

XIII.5.3 Expansion of the Velocity Approximation

A similar expansion holds also for the velocities. We show this for the scheme (2.11) or its equivalent one-step formulation (2.8) with (2.9).

Theorem 5.3. *Under the assumptions of Theorem 5.2, the velocity approximation \dot{x}_n given by (2.11) has an expansion*

$$\dot{x}_n = v_h(t) + \sum_{0 < |k| < N} e^{ik\omega t} w_h^k(t) + \mathcal{O}(t^2h^{N-1})$$

uniformly for $0 \leq t = nh \leq T$, where the real-valued functions $v_h = (v_{h,0}, v_{h,1})$ and the complex-valued functions $w_h^k = (w_{h,0}^k, w_{h,1}^k)$ together with their derivatives up to arbitrary order satisfy

$$\begin{aligned}
v_{h,0} &= \dot{y}_{h,0} + \mathcal{O}(h^2), & w_{h,0}^1 &= \mathcal{O}(\omega^{-1}), & w_{h,0}^k &= \mathcal{O}(\omega^{-k}) \\
w_{h,1}^1 &= i\omega z_{h,1}^1 + \mathcal{O}(\omega^{-1}), & v_{h,1} &= \mathcal{O}(\omega^{-1}), & w_{h,1}^k &= \mathcal{O}(\omega^{-k})
\end{aligned} \tag{5.23}$$

for $k = 2, \dots, N-1$. Moreover, $w_h^{-k} = \overline{w_h^k}$. The constants symbolized by the \mathcal{O} -notation are independent of ω and h with (5.10), but depend on E , N , c , and T .

Proof. Let $u_h(t)$ be defined by the continuous analogue of (2.11),

$$2h \operatorname{sinc}(h\Omega) u_h(t) = x_h(t+h) - x_h(t-h). \tag{5.24}$$

Theorem 5.2 then yields that

$$\dot{x}_n = u_h(t) + \mathcal{O}(t^2h^{N-1})$$

for $t = nh$ on bounded time intervals. Here we used that the remainder term in the lower component of (5.12) is of the form $\mathcal{O}(\psi(h\omega)(\phi(h\omega) + h)t^2h^N)$, so that its

quotient with $2h \operatorname{sinc}(h\omega)$ becomes $\mathcal{O}(t^2 h^{N-1})$ by the third of the conditions (4.1) and by (5.10). The function $u_h(t)$ can be written as

$$u_h(t) = v_h(t) + \sum_{0 < |k| < N} e^{ik\omega t} w_h^k(t). \quad (5.25)$$

We insert the relation (5.14) into $-i \sin(ihD)x_h(t) = h \operatorname{sinc}(h\Omega)u_h(t)$, which is equivalent to (5.24), and compare the coefficients of $e^{ik\omega t}$ to obtain

$$\begin{aligned} \operatorname{sinc}(ihD) \dot{y}_{h,0} &= v_{h,0} \\ \operatorname{sinc}(ihD) \dot{y}_{h,1} &= \operatorname{sinc}(h\omega) v_{h,1} \\ (ih)^{-1} \sin(ihD - kh\omega) z_{h,0}^k &= w_{h,0}^k \\ (ih)^{-1} \sin(ihD - kh\omega) z_{h,1}^k &= \operatorname{sinc}(h\omega) w_{h,1}^k \end{aligned} \quad (5.26)$$

for $k = 1, \dots, N-1$. In particular, for $w_{h,1}^1$ we get

$$w_{h,1}^1 = i\omega \cos(ihD) z_{h,1}^1 - i\omega \frac{\cos(h\omega)}{\sin(h\omega)} \sin(ihD) z_{h,1}^1. \quad (5.27)$$

With the above equations, the estimates now follow with the bounds (5.21) of the coefficient functions and their derivatives, using again (4.1). \square

XIII.6 Almost-Invariants of the Modulated Fourier Expansions

The system for the coefficients of the modulated Fourier expansion of the exact solution is shown to have two formal invariants, which are related to the total and the oscillatory energy. In particular, this explains the near-conservation of the oscillatory energy over very long times. Analogous almost-invariants are shown to exist also for the modulated Fourier expansion of the numerical solution. This forms the basis for results on the long-time energy conservation of numerical methods, which will be given in Sections XIII.7 and XIII.8.

XIII.6.1 The Hamiltonian of the Modulated Fourier Expansion

The equation (2.1) is a Hamiltonian system with the Hamiltonian

$$H(x, \dot{x}) = \frac{1}{2} \dot{x}^T \dot{x} + \frac{1}{2} x^T \Omega^2 x + U(x). \quad (6.1)$$

In the modulated Fourier expansion of the solution $x(t)$ of (2.1), denote $y^0(t) = y(t)$ and $y^k(t) = e^{ik\omega t} z^k(t)$ ($0 < |k| < N$), and let

$$\mathbf{y} = (y^{-N+1}, \dots, y^{-1}, y^0, y^1, \dots, y^{N-1}).$$

By (5.5)–(5.7) these functions satisfy

$$\ddot{y}^k + \Omega^2 y^k = - \sum_{s(\alpha)=k} \frac{1}{m!} U^{(m+1)}(y^0) \mathbf{y}^\alpha + \mathcal{O}(\omega^{-N}). \quad (6.2)$$

Here, the sum is over all $m \geq 0$ and all multi-indices $\alpha = (\alpha_1, \dots, \alpha_m)$ with integers α_j ($0 < |\alpha_j| < N$) which have a given sum $s(\alpha) = \sum_{j=1}^m \alpha_j$, and we write $\mathbf{y}^\alpha = (y^{\alpha_1}, \dots, y^{\alpha_m})$. We define

$$\mathcal{U}(\mathbf{y}) = U(y^0) + \sum_{s(\alpha)=0} \frac{1}{m!} U^{(m)}(y^0) \mathbf{y}^\alpha. \quad (6.3)$$

From the above it follows that $\mathbf{y}(t)$ satisfies the system

$$\ddot{y}^k + \Omega^2 y^k = - \nabla_{y^{-k}} \mathcal{U}(\mathbf{y}) + \mathcal{O}(\omega^{-N}) \quad (6.4)$$

which, neglecting the $\mathcal{O}(\omega^{-N})$ term, is the Hamiltonian system (cf. Exercise 6)

$$\dot{y}^k = \frac{\partial \mathcal{H}}{\partial \dot{y}^{-k}}(\mathbf{y}, \dot{\mathbf{y}}), \quad \ddot{y}^k = - \frac{\partial \mathcal{H}}{\partial y^{-k}}(\mathbf{y}, \dot{\mathbf{y}}) \quad (6.5)$$

with

$$\mathcal{H}(\mathbf{y}, \dot{\mathbf{y}}) = \frac{1}{2} \sum_{|k| < N} \left((\dot{y}^{-k})^T \dot{y}^k + (y^{-k})^T \Omega^2 y^k \right) + \mathcal{U}(\mathbf{y}). \quad (6.6)$$

Theorem 6.1. *Under the assumptions of Theorem 5.1, the Hamiltonian of the modulated Fourier expansion satisfies*

$$\mathcal{H}(\mathbf{y}(t), \dot{\mathbf{y}}(t)) = \mathcal{H}(\mathbf{y}(0), \dot{\mathbf{y}}(0)) + \mathcal{O}(\omega^{-N}) \quad (6.7)$$

$$\mathcal{H}(\mathbf{y}(t), \dot{\mathbf{y}}(t)) = H(x(t), \dot{x}(t)) + \mathcal{O}(\omega^{-1}). \quad (6.8)$$

The constants symbolized by \mathcal{O} are independent of ω and t with $0 \leq t \leq T$, but depend on E , N and T .

Proof. Multiplying (6.4) with $(\dot{y}^{-k})^T$ and summing up gives

$$\sum_{|k| < N} (\dot{y}^{-k})^T (\ddot{y}^k + \Omega^2 y^k) = - \frac{d}{dt} \mathcal{U}(\mathbf{y}) + \mathcal{O}(\omega^{-N}). \quad (6.9)$$

Integrating from 0 to t and using $y^{-k} = \overline{y^k}$ then yields (6.7).

By the bounds of Theorem 5.1, we have for $0 \leq t \leq T$

$$\mathcal{H}(\mathbf{y}, \dot{\mathbf{y}}) = \frac{1}{2} \|\dot{y}_0^0\|^2 + \|\dot{y}_1^1\|^2 + \omega^2 \|y_1^1\|^2 + U(y^0) + \mathcal{O}(\omega^{-1}). \quad (6.10)$$

On the other hand, we have from (6.1) and (5.1)

$$H(x, \dot{x}) = \frac{1}{2} \|\dot{y}_0^0\|^2 + \frac{1}{2} \|\dot{y}_1^1 + \dot{y}_1^{-1}\|^2 + \frac{1}{2} \omega^2 \|y_1^1 + y_1^{-1}\|^2 + U(y^0) + \mathcal{O}(\omega^{-1}). \quad (6.11)$$

Using $y_1^1 = e^{i\omega t} z_1^1$ and $\dot{y}_1^1 = e^{i\omega t} (\dot{z}_1^1 + i\omega z_1^1)$ together with $y_1^{-1} = \overline{y_1^1}$, it follows from $\dot{z}_1^1 = \mathcal{O}(\omega^{-1})$ that $\dot{y}_1^1 + \dot{y}_1^{-1} = i\omega(y_1^1 - y_1^{-1}) + \mathcal{O}(\omega^{-1})$ and $\|\dot{y}_1^1\| = \omega \|y_1^1\| + \mathcal{O}(\omega^{-1})$. Inserted into (6.10) and (6.11), this yields (6.8). \square

XIII.6.2 A Formal Invariant Close to the Oscillatory Energy

In addition to the Hamiltonian $\mathcal{H}(\mathbf{y}, \dot{\mathbf{y}})$, the system for the coefficients of the modulated Fourier expansion has another formally conserved quantity. This almost-invariant depends only on the oscillating part and is given by

$$\mathcal{I}(\mathbf{y}, \dot{\mathbf{y}}) = -i\omega \sum_{0 < |k| < N} k (y^{-k})^T \dot{y}^k. \quad (6.12)$$

This turns out to be close to the energy of the harmonic oscillator,

$$I(x, \dot{x}) = \frac{1}{2} \|\dot{x}_1\|^2 + \frac{1}{2} \omega^2 \|x_1\|^2. \quad (6.13)$$

Theorem 6.2. *Under the assumptions of Theorem 5.1,*

$$\mathcal{I}(\mathbf{y}(t), \dot{\mathbf{y}}(t)) = \mathcal{I}(\mathbf{y}(0), \dot{\mathbf{y}}(0)) + \mathcal{O}(\omega^{-N}) \quad (6.14)$$

$$\mathcal{I}(\mathbf{y}(t), \dot{\mathbf{y}}(t)) = I(x(t), \dot{x}(t)) + \mathcal{O}(\omega^{-1}). \quad (6.15)$$

The constants symbolized by \mathcal{O} are independent of ω and t with $0 \leq t \leq T$, but depend on E , N and T .

Proof. For the vector

$$\mathbf{y}(\lambda) = (e^{i(-N+1)\lambda} y^{-N+1}, \dots, e^{-i\lambda} y^{-1}, y^0, e^{i\lambda} y^1, \dots, e^{i(N-1)\lambda} y^{N-1})$$

the definition (6.3) of \mathcal{U} shows that $\mathcal{U}(\mathbf{y}(\lambda))$ does not depend on λ . Its derivative with respect to λ thus yields

$$0 = \frac{d}{d\lambda} \mathcal{U}(\mathbf{y}(\lambda)) = \sum_{0 < |k| < N} ik e^{ik\lambda} (y^k)^T \nabla_k \mathcal{U}(\mathbf{y}(\lambda)),$$

and putting $\lambda = 0$ we obtain

$$\sum_{0 < |k| < N} ik (y^k)^T \nabla_k \mathcal{U}(\mathbf{y}) = 0 \quad (6.16)$$

for all vectors $\mathbf{y} = (y^{-N+1}, \dots, y^{-1}, y^0, y^1, \dots, y^{N-1})$.

The proof of Theorem 6.2 is now very similar to that of Theorem 6.1. We multiply the relation (6.4) with $-i\omega k (y^{-k})^T$ instead of $(\dot{y}^{-k})^T$. Summing up yields, with the use of (6.16),

$$-i\omega \sum_{0 < |k| < N} k (y^{-k})^T (\ddot{y}^k + \Omega^2 y^k) = \mathcal{O}(\omega^{-N}). \quad (6.17)$$

The time derivative of $\mathcal{I}(\mathbf{y}, \dot{\mathbf{y}})$ of (6.12) equals

$$\frac{d}{dt} \mathcal{I}(\mathbf{y}, \dot{\mathbf{y}}) = -i\omega \sum_{0 < |k| < N} k \left((y^{-k})^T \ddot{y}^k + (\dot{y}^{-k})^T \dot{y}^k \right). \quad (6.18)$$

In the sums $\sum_k k(y^{-k})^T \Omega^2 y^k$ and $\sum_k k(\dot{y}^{-k})^T \dot{y}^k$, the terms with k and $-k$ cancel. Hence, (6.17) and (6.18) together yield

$$\frac{d}{dt} \mathcal{I}(\mathbf{y}, \dot{\mathbf{y}}) = \mathcal{O}(\omega^{-N}),$$

which implies (6.14).

With $\dot{y}^k = e^{ik\omega t}(z^k + ik\omega z^k) = ik\omega y^k + \mathcal{O}(\omega^{-1})$, it follows from the bounds of Theorem 5.1 that

$$\mathcal{I}(\mathbf{y}, \dot{\mathbf{y}}) = 2\omega^2 \|y_1^1\|^2 + \mathcal{O}(\omega^{-1}).$$

On the other hand, using the arguments of the proof of Theorem 6.1, we have

$$I(x, \dot{x}) = \frac{1}{2} \|\dot{y}_1^1 + \dot{y}_1^{-1}\|^2 + \frac{1}{2} \omega^2 \|y_1^1 + y_1^{-1}\|^2 + \mathcal{O}(\omega^{-1}) = 2\omega^2 \|y_1^1\|^2 + \mathcal{O}(\omega^{-1}).$$

This proves the second statement of the theorem. \square

Theorem 6.2 implies that the oscillatory energy is nearly conserved over long times:

Theorem 6.3. *If the solution $x(t)$ of (2.1) stays in a compact set for $0 \leq t \leq \omega^N$, then*

$$I(x(t), \dot{x}(t)) = I(x(0), \dot{x}(0)) + \mathcal{O}(\omega^{-1}) + \mathcal{O}(t\omega^{-N}).$$

The constants symbolized by \mathcal{O} are independent of ω and t with $0 \leq t \leq \omega^N$, but depend on E and N .

Proof. With a fixed $T > 0$, let \mathbf{y}_j denote the vector of the modulated Fourier expansion terms that correspond to starting values $(x(jT), \dot{x}(jT))$. For $t = (n + \theta)T$ with $0 \leq \theta < 1$, we have by (6.15)

$$\begin{aligned} I(x(t), \dot{x}(t)) - I(x(0), \dot{x}(0)) &= \mathcal{I}(\mathbf{y}_n(\theta T), \dot{\mathbf{y}}_n(\theta T)) + \mathcal{O}(\omega^{-1}) - \mathcal{I}(\mathbf{y}_0(0), \dot{\mathbf{y}}_0(0)) + \mathcal{O}(\omega^{-1}) \\ &= \mathcal{I}(\mathbf{y}_n(\theta T), \dot{\mathbf{y}}_n(\theta T)) - \mathcal{I}(\mathbf{y}_n(0), \dot{\mathbf{y}}_n(0)) + \\ &\quad \sum_{j=0}^{n-1} \left(\mathcal{I}(\mathbf{y}_{j+1}(0), \dot{\mathbf{y}}_{j+1}(0)) - \mathcal{I}(\mathbf{y}_j(0), \dot{\mathbf{y}}_j(0)) \right) + \mathcal{O}(\omega^{-1}). \end{aligned}$$

We note

$$\mathcal{I}(\mathbf{y}_{j+1}(0), \dot{\mathbf{y}}_{j+1}(0)) - \mathcal{I}(\mathbf{y}_j(0), \dot{\mathbf{y}}_j(0)) = \mathcal{O}(\omega^{-N}),$$

because, by the quasi-uniqueness of the coefficient functions as stated by Theorem 5.1, we have $\mathbf{y}_{j+1}(0) = \mathbf{y}_j(T) + \mathcal{O}(\omega^{-N})$ and $\dot{\mathbf{y}}_{j+1}(0) = \dot{\mathbf{y}}_j(T) + \mathcal{O}(\omega^{-N})$, and we have the bound (6.14) of Theorem 6.2. The same argument applies to $\mathcal{I}(\mathbf{y}_n(\theta T), \dot{\mathbf{y}}_n(\theta T)) - \mathcal{I}(\mathbf{y}_n(0), \dot{\mathbf{y}}_n(0))$. This yields the result. \square

In a different approach, Benettin, Galgani & Giorgilli (1987) use a sequence of coordinate transformations from Hamiltonian perturbation theory to show that I has only small deviations over time intervals which grow exponentially with ω , in the case of an analytic potential U . By carefully tracing the dependence on N of the constants in the $\mathcal{O}(\omega^{-N})$ -terms, near-conservation of I over exponentially long time intervals can be shown also within the present framework of modulated Fourier expansions; see Cohen, Hairer & Lubich (2003).

XIII.6.3 Almost-Invariants of the Numerical Method

We show that the coefficients of the modulated Fourier expansion of the numerical solution have almost-invariants that are obtained similarly to the above. We denote

$$\begin{aligned}\mathbf{y}_h &= (y_h^{-N+1}, \dots, y_h^{-1}, y_h^0, y_h^1, \dots, y_h^{N-1}) \\ \mathbf{z}_h &= (z_h^{-N+1}, \dots, z_h^{-1}, z_h^0, z_h^1, \dots, z_h^{N-1})\end{aligned}$$

with $y_h^0(t) = z_h^0(t) = y_h(t)$ and $y_h^k(t) = e^{ik\omega t} z_h^k(t)$, where y_h and z_h^k are the coefficients of the modulated Fourier expansion of Theorem 5.2. Similar to (6.3) we consider the function

$$\mathcal{U}_h(\mathbf{y}_h) = U(\Phi y_h^0) + \sum_{s(\alpha)=0} \frac{1}{m!} U^{(m)}(\Phi y_h^0)(\Phi \mathbf{y}_h)^\alpha, \quad (6.19)$$

where the sum is again taken over all $m \geq 1$ and all multi-indices $\alpha = (\alpha_1, \dots, \alpha_m)$ with $0 < |\alpha_j| < N$ for which $s(\alpha) = \sum_j \alpha_j = 0$. It then follows from (5.22), multiplied with $h^{-2}\Psi^{-1}\Phi$, that the functions $y_h^k(t)$ satisfy

$$\Psi^{-1}\Phi h^{-2}\mathcal{L}(hD)y_h^k = -\nabla_{-k}\mathcal{U}_h(\mathbf{y}_h) + \mathcal{O}(h^N), \quad (6.20)$$

where $\mathcal{L}(hD)$ of (5.16) denotes again the difference operator of the numerical method. The similarity of these relations to (6.4) allows us to obtain almost-conserved quantities that are analogues of \mathcal{H} and \mathcal{I} above.

The First Almost-Invariant. We multiply (6.20) by $(\dot{y}_h^{-k})^T$, and as in (6.9) we obtain

$$\sum_{|k|<N} (\dot{y}_h^{-k})^T \Psi^{-1}\Phi h^{-2}\mathcal{L}(hD)y_h^k + \frac{d}{dt}\mathcal{U}_h(\mathbf{y}_h) = \mathcal{O}(h^N).$$

Since we know bounds of the coefficient functions z_h^k and of their derivatives from Theorem 5.2, we switch to the quantities z_h^k and we get the equivalent relation

$$\sum_{|k|<N} (\dot{z}_h^{-k} - ik\omega z_h^{-k})^T \Psi^{-1}\Phi h^{-2}\mathcal{L}(hD + ik\omega h)z_h^k + \frac{d}{dt}\mathcal{U}_h(\mathbf{z}_h) = \mathcal{O}(h^N). \quad (6.21)$$

We shall show that the left-hand side is the total derivative of an expression that depends only on z_h^k and derivatives thereof. Consider first the term for $k = 0$. The

symmetry of the numerical method enters at this very point in the way that the expression $\mathcal{L}(hD)y = h^2\ddot{y} + c_4h^4y^{(4)} + c_6h^6y^{(6)} + \dots$ contains only terms with derivatives of an even order. Multiplied with \dot{y}^T , even-order derivatives of y give a total derivative:

$$\dot{y}^T y^{(2l)} = \frac{d}{dt} \left(\dot{y}^T y^{(2l-1)} - \ddot{y}^T y^{(2l-2)} + \dots \mp (y^{(l-1)})^T y^{(l+1)} \pm \frac{1}{2} (y^{(l)})^T y^{(l)} \right).$$

Thanks to the symmetry of the difference operator $\mathcal{L}(hD)$ only expressions of this type appear in the term for $k = 0$ in (6.21), with z_h^0 in the role of y . Similarly, we get for $z = z_h^k$ and $\bar{z} = z_h^{-k}$ with $0 < |k| < N$

$$\begin{aligned} \operatorname{Re} \dot{\bar{z}}^T z^{(2l)} &= \operatorname{Re} \frac{d}{dt} \left(\dot{\bar{z}}^T z^{(2l-1)} - \dots \mp (\bar{z}^{(l-1)})^T z^{(l+1)} \pm \frac{1}{2} (\bar{z}^{(l)})^T z^{(l)} \right) \\ \operatorname{Re} \bar{z}^T z^{(2l+1)} &= \operatorname{Re} \frac{d}{dt} \left(\bar{z}^T z^{(2l)} - \dots \pm (\bar{z}^{(l-1)})^T z^{(l+1)} \mp \frac{1}{2} (\bar{z}^{(l)})^T z^{(l)} \right) \\ \operatorname{Im} \dot{\bar{z}}^T z^{(2l+1)} &= \operatorname{Im} \frac{d}{dt} \left(\dot{\bar{z}}^T z^{(2l)} - \ddot{\bar{z}}^T z^{(2l-1)} + \dots \mp (\bar{z}^{(l)})^T z^{(l+1)} \right) \\ \operatorname{Im} \bar{z}^T z^{(2l+2)} &= \operatorname{Im} \frac{d}{dt} \left(\bar{z}^T z^{(2l+1)} - \dot{\bar{z}}^T z^{(2l)} + \dots \pm (\bar{z}^{(l)})^T z^{(l+1)} \right). \end{aligned}$$

Using the formulas (5.19) for $\mathcal{L}(hD + ikh\omega)$, it is seen that the term for k in (6.21) has an asymptotic h -expansion with expressions of the above type as coefficients. The left-hand side of (6.21) can therefore be written as the time derivative of a function $\widehat{\mathcal{H}}_h[\mathbf{z}_h](t)$ which depends on the values at t of the coefficient function vector \mathbf{z}_h and its first N time derivatives. The relation (6.21) thus becomes

$$\frac{d}{dt} \widehat{\mathcal{H}}_h[\mathbf{z}_h](t) = \mathcal{O}(h^N).$$

Together with the estimates of Theorem 5.2, this construction of $\widehat{\mathcal{H}}_h$ yields the following result.

Lemma 6.4. *Under the assumptions of Theorem 5.2, the coefficient functions $\mathbf{z}_h = (z_h^{-N+1}, \dots, z_h^{-1}, y_h, z_h^1, \dots, z_h^{N-1})$ of the modulated Fourier expansion of the numerical solution satisfy*

$$\widehat{\mathcal{H}}_h[\mathbf{z}_h](t) = \widehat{\mathcal{H}}_h[\mathbf{z}_h](0) + \mathcal{O}(th^N) \quad (6.22)$$

for $0 \leq t \leq T$. Moreover,

$$\widehat{\mathcal{H}}_h[\mathbf{z}_h](t) = \frac{1}{2} \|\dot{y}_{h,0}(t)\|^2 + \sigma(h\omega) 2\omega^2 \|z_{h,1}^1(t)\|^2 + U(\Phi y_h(t)) + \mathcal{O}(h^2), \quad (6.23)$$

where $\sigma(h\omega) = \operatorname{sinc}(h\omega)\phi(h\omega)/\psi(h\omega)$. \square

The Second Almost-Invariant. By the same calculation as in the proof of Theorem 6.2 we obtain for $\mathcal{U}_h(\mathbf{y}_h(t))$ of (6.19)

$$0 = \sum_{0 < |k| < N} ik\omega(y_h^k)^T \nabla_k \mathcal{U}_h(\mathbf{y}_h) .$$

It then follows from (6.20) that

$$-i\omega \sum_{0 < |k| < N} k(y_h^{-k})^T \Psi^{-1} \Phi h^{-2} \mathcal{L}(hD) y_h^k = \mathcal{O}(h^N) .$$

Written in the z variables, this becomes

$$-i\omega \sum_{0 < |k| < N} k(z_h^{-k})^T \Psi^{-1} \Phi h^{-2} \mathcal{L}(hD + ik\omega h) z_h^k = \mathcal{O}(h^N) . \quad (6.24)$$

As in (6.21), the left-hand expression can be written as the time derivative of a function $\widehat{\mathcal{I}}_h[\mathbf{z}_h](t)$ which depends on the values at t of the function \mathbf{z}_h and its first N derivatives:

$$\frac{d}{dt} \widehat{\mathcal{I}}_h[\mathbf{z}_h](t) = \mathcal{O}(h^N) .$$

Together with the estimates of Theorem 5.2 this yields the following result.

Lemma 6.5. *Under the assumptions of Theorem 5.2, the coefficient functions \mathbf{z}_h of the modulated Fourier expansion of the numerical solution satisfy*

$$\widehat{\mathcal{I}}_h[\mathbf{z}_h](t) = \widehat{\mathcal{I}}_h[\mathbf{z}_h](0) + \mathcal{O}(th^N) \quad (6.25)$$

for $0 \leq t \leq T$. Moreover,

$$\widehat{\mathcal{I}}_h[\mathbf{z}_h](t) = \sigma(h\omega) 2\omega^2 \|z_{h,1}^1(t)\|^2 + \mathcal{O}(h^2) , \quad (6.26)$$

where again $\sigma(h\omega) = \text{sinc}(h\omega)\phi(h\omega)/\psi(h\omega)$. \square

Symplectic methods have $\psi(\xi) = \text{sinc}(\xi)\phi(\xi)$ and hence $\sigma(h\omega) = 1$. To be able to also treat methods where $\sigma(h\omega)$ can be small, we need to sharpen the estimates of Lemma 6.5. Close scrutiny of the equations (5.20) that determine the coefficient functions of the modulated Fourier expansion, shows that the $\mathcal{O}(h^2)$ term in (6.26) contains a factor $\phi(h\omega)^2$, and that the $\mathcal{O}(th^N)$ term in (6.25) can be put in the form $\mathcal{O}(t\phi(h\omega)^N h^N) + \mathcal{O}(th^{N+m})$ with an arbitrary integer $m \geq 0$; cf. (5.12). Assume now that

$$\phi \text{ is analytic with no real zeros other than integral multiples of } \pi. \quad (6.27)$$

This condition ensures that $|\phi(h\omega)|^2 \geq ch^m$ for some m if $h\omega$ satisfies (5.10). Under the conditions of Theorem 5.2, in particular, (4.1) and (5.10), the improved bounds of the remainder terms yield the following estimates for $\mathcal{I}_h = \widehat{\mathcal{I}}_h/\sigma(h\omega)$:

$$\mathcal{I}_h[\mathbf{z}_h](t) = \mathcal{I}_h[\mathbf{z}_h](0) + \mathcal{O}(th^N) \quad (6.28)$$

$$\mathcal{I}_h[\mathbf{z}_h](t) = 2\omega^2 \|z_{h,1}^1(t)\|^2 + \mathcal{O}(h^2) . \quad (6.29)$$

Relationship with the Total and the Oscillatory Energy. The almost-invariants

$$\mathcal{I}_h = \frac{1}{\sigma(h\omega)} \widehat{\mathcal{I}}_h, \quad \mathcal{H}_h = \widehat{\mathcal{H}}_h - \left(1 - \frac{1}{\sigma(h\omega)}\right) \widehat{\mathcal{I}}_h \quad (6.30)$$

of the coefficient functions of the modulated Fourier expansion are then close to the total energy H and the oscillatory energy I along the numerical solution (x_n, \dot{x}_n) :

Theorem 6.6. *Under the conditions of Theorems 5.2 and condition (6.27),*

$$\begin{aligned} \mathcal{H}_h[\mathbf{z}_h](t) &= \mathcal{H}_h[\mathbf{z}_h](0) + \mathcal{O}(th^N), & \mathcal{I}_h[\mathbf{z}_h](t) &= \mathcal{I}_h[\mathbf{z}_h](0) + \mathcal{O}(th^N) \\ \mathcal{H}_h[\mathbf{z}_h](t) &= H(x_n, \dot{x}_n) + \mathcal{O}(h), & \mathcal{I}_h[\mathbf{z}_h](t) &= I(x_n, \dot{x}_n) + \mathcal{O}(h) \end{aligned}$$

holds for $0 \leq t = nh \leq T$. The constants symbolized by \mathcal{O} depend on E , N and T .

Proof. The upper two relations follow directly from (6.22) and (6.28). Theorems 5.2 and 5.3 show

$$\begin{aligned} \omega x_{n,1} &= \omega(e^{i\omega t} z_{h,1}^1(t) + e^{-i\omega t} z_{h,1}^{-1}(t)) + \mathcal{O}(h) \\ \dot{x}_{n,1} &= i\omega(e^{i\omega t} z_{h,1}^1(t) - e^{-i\omega t} z_{h,1}^{-1}(t)) + \mathcal{O}(h) . \end{aligned}$$

With the identity $\|v + \bar{v}\|^2 + \|v - \bar{v}\|^2 = 4\|v\|^2$, this implies

$$I(x_n, \dot{x}_n) = 2\omega^2 \|z_{h,1}^1(t)\|^2 + \mathcal{O}(h) .$$

A comparison with (6.29) then gives the stated relation between I and \mathcal{I}_h . The relation between H and \mathcal{H}_h is proved in the same way, using in addition (6.23). \square

XIII.7 Long-Time Near-Conservation of Total and Oscillatory Energy

With the results of the previous section, we can now show that the numerical method nearly preserves the total energy H and the oscillatory energy I over time intervals of length $C_N h^{-N+1}$, for any N for which the non-resonance condition (5.10) is satisfied. Such a result is due to Hairer & Lubich (2000a).

For convenience we restate the assumptions:

- the energy bound (2.3): $\frac{1}{2}\|\dot{x}(0)\|^2 + \frac{1}{2}\|\Omega x(0)\|^2 \leq E$;
- the condition on the numerical solution: the values Φx_n stay in a compact subset of a domain on which the potential U is smooth;

- the conditions on the filter functions: ψ and ϕ are even, real-analytic, and have no real zeros other than integral multiples of π ; they satisfy $\psi(0) = \phi(0) = 1$ and (4.1):

$$\begin{aligned} |\psi(h\omega)| &\leq C_1 \operatorname{sinc}^2(\tfrac{1}{2}h\omega), & |\phi(h\omega)| &\leq C_2 |\operatorname{sinc}(\tfrac{1}{2}h\omega)|, \\ |\psi(h\omega)\phi(h\omega)| &\leq C_3 |\operatorname{sinc}(h\omega)|; \end{aligned} \quad (7.1)$$

- the condition $h\omega \geq c_0 > 0$;
- the non-resonance condition (5.10): for some $N \geq 2$,

$$|\sin(\tfrac{1}{2}kh\omega)| \geq c\sqrt{h} \quad \text{for } k = 1, \dots, N.$$

Theorem 7.1. *Under the above conditions, the numerical solution of (2.1) obtained by the method (2.7)–(2.8) with (2.9) satisfies*

$$\begin{aligned} H(x_n, \dot{x}_n) &= H(x_0, \dot{x}_0) + \mathcal{O}(h) \\ I(x_n, \dot{x}_n) &= I(x_0, \dot{x}_0) + \mathcal{O}(h) \end{aligned} \quad \text{for } 0 \leq nh \leq h^{-N+1}.$$

The constants symbolized by \mathcal{O} are independent of n, h, ω satisfying the above conditions, but depend on N and the constants in the conditions.

Proof. The estimates of Theorem 6.6 hold uniformly over bounded intervals. We now apply those estimates repeatedly on intervals of length h , for modulated Fourier expansions corresponding to different starting values. As long as (x_n, \dot{x}_n) satisfies the bounded-energy condition (2.3) (possibly with a larger constant E), Theorem 5.2 gives a modulated Fourier expansion that corresponds to starting values (x_n, \dot{x}_n) . We denote the vector of coefficient functions of this expansion by $\mathbf{z}_n(t)$:

$$\mathbf{z}_n = (z_n^{-N+1}, \dots, z_n^{-1}, y_n, z_n^1, \dots, z_n^{N-1})$$

(omitting the notational dependence on h for simplicity). Because of the uniqueness, up to $\mathcal{O}(h^{N+1})$, of the coefficient functions of the modulated Fourier expansion constructed by (5.20), the following diagram commutes up to terms of size $\mathcal{O}(h^{N+1})$:

$$\begin{array}{ccc} (x_n, \dot{x}_n) & \longleftrightarrow & (\mathbf{z}_n(0), \dot{\mathbf{z}}_n(0)) \\ & & \downarrow \text{flow} \\ \downarrow \text{numerical} & & (\mathbf{z}_n(h), \dot{\mathbf{z}}_n(h)) \\ \text{method} & & = (\text{up to } \mathcal{O}(h^{N+1})) \\ (x_{n+1}, \dot{x}_{n+1}) & \longleftrightarrow & (\mathbf{z}_{n+1}(0), \dot{\mathbf{z}}_{n+1}(0)) \end{array}$$

The construction of the coefficient functions via (5.20) shows that also higher derivatives of \mathbf{z}_n at h and \mathbf{z}_{n+1} at 0 differ by only $\mathcal{O}(h^{N+1})$. From the above diagram and Theorem 6.6 we thus obtain

$$\begin{aligned}\mathcal{H}_h[\mathbf{z}_{n+1}](0) &= \mathcal{H}_h[\mathbf{z}_n](h) + \mathcal{O}(h^{N+1}) \\ &= \mathcal{H}_h[\mathbf{z}_n](0) + \mathcal{O}(h^{N+1}).\end{aligned}$$

Repeated use of this relation gives

$$\mathcal{H}_h[\mathbf{z}_n](0) = \mathcal{H}_h[\mathbf{z}_0](0) + \mathcal{O}(nh^{N+1}).$$

Moreover, by Theorem 6.6 the coefficient functions corresponding to the starting values (x_n, \dot{x}_n) and (x_0, \dot{x}_0) satisfy

$$\begin{aligned}\mathcal{H}_h[\mathbf{z}_n](0) &= H(x_n, \dot{x}_n) + \mathcal{O}(h), \\ \mathcal{H}_h[\mathbf{z}_0](0) &= H(x_0, \dot{x}_0) + \mathcal{O}(h).\end{aligned}$$

So we obtain

$$\begin{aligned}H(x_n, \dot{x}_n) - H(x_0, \dot{x}_0) &= \mathcal{H}_h[\mathbf{z}_n](0) - \mathcal{H}_h[\mathbf{z}_0](0) + \mathcal{O}(h) \\ &= \mathcal{O}(nh^{N+1}) + \mathcal{O}(h),\end{aligned}$$

which gives the desired bound for the deviation of the total energy along the numerical solution. The same argument applies to $I(x_n, \dot{x}_n)$. \square

The imposed bounds of ψ and ϕ become important when $h\omega$ is close to an integral multiple of π . Are these conditions also sufficient to guarantee favourable energy behaviour uniformly in $h\omega$, arbitrarily close to multiples of π ? Unfortunately the answer is negative (see Fig. 2.5 to Fig. 2.7). The analysis of method (2.7)–(2.9) for exact resonances $h\omega = m\pi$ with integer m shows that stronger conditions

$$|\psi(h\omega)| \leq C |\operatorname{sinc}(h\omega)|, \quad |\psi(h\omega)\phi(h\omega)| \leq C \operatorname{sinc}^2(h\omega) \quad (7.2)$$

are required. Even this is not sufficient for near-conservation of the total and the oscillatory energy for $h\omega$ near a multiple of π . For linear problems

$$\ddot{x} + \begin{pmatrix} 0 & 0 \\ 0 & \omega^2 \end{pmatrix} x = -Ax$$

with a two-dimensional symmetric matrix A with $a_{00} > 0$, and with initial values satisfying the bounded-energy condition (2.3), Hairer & Lubich (2000a) show that the numerical method conserves the total energy up to $\mathcal{O}(h)$ uniformly for all times and for all values of $h\omega$, if and only if

$$\psi(\xi) = \operatorname{sinc}^2(\xi) \phi(\xi). \quad (7.3)$$

There is *no* method (2.7)–(2.8) which approximately preserves the oscillatory energy I uniformly for all $h\omega$ in a fixed open interval that contains a multiple of 2π .

In summary, the bad effect of step-size resonances on the energy behaviour of the method cannot be eliminated, but it can be considerably mitigated by an appropriate choice of the filter functions ψ and ϕ .

XIII.8 Energy Behaviour of the Störmer–Verlet Method

The results of Sections XIII.5–XIII.7 provide new insight into the energy behaviour of the classical Störmer–Verlet method. We present in this section weakened versions of results of Hairer & Lubich (2000b).

In applications, the Störmer–Verlet method is typically used with step sizes h for which the product with the highest frequency ω is in the range of linear stability, but is bounded away from 0. For example, in spatially discretized wave equations, $h\omega$ is known as the CFL number, which is typically kept near 1. Values of $h\omega$ around $\frac{1}{2}$ are often used in molecular dynamics. In contrast, the backward error analysis of Chap. IX explains the long-time energy behaviour only for $h\omega \rightarrow 0$.

Consider now applying the Störmer–Verlet method to the nonlinear model problem (2.1)–(2.3),

$$x_{n+1} - 2x_n + x_{n-1} = -h^2\Omega^2x_n - h^2\nabla U(x_n), \quad (8.1)$$

with $h\omega < 2$ for linear stability. The method is made accessible to the analysis of Sections XIII.3–XIII.7 by rewriting it as a trigonometric method (2.6) with a *modified frequency*:

$$x_{n+1} - 2\cos(h\tilde{\Omega})x_n + x_{n-1} = -h^2\nabla U(x_n), \quad (8.2)$$

where

$$\tilde{\Omega} = \begin{pmatrix} 0 & 0 \\ 0 & \tilde{\omega}I \end{pmatrix} \quad \text{with} \quad \sin(\tfrac{1}{2}h\tilde{\omega}) = \tfrac{1}{2}h\omega. \quad (8.3)$$

The velocity approximation

$$\dot{x}_n = \frac{x_{n+1} - x_{n-1}}{2h}$$

does not correspond to the velocity approximation (2.11) of the trigonometric method, but this presents only a minor technical difficulty. We show that the following *modified energies* are well conserved by the Störmer–Verlet method:

$$\begin{aligned} H^*(x, \dot{x}) &= H(x, \dot{x}) + \tfrac{1}{2}\gamma \|\dot{x}_1\|^2 \\ I^*(x, \dot{x}) &= I(x, \dot{x}) + \tfrac{1}{2}\gamma \|\dot{x}_1\|^2 \end{aligned} \quad \text{with} \quad \gamma = \frac{1}{1 - \frac{1}{4}(h\omega)^2} - 1. \quad (8.4)$$

Here H and I are again the total and the oscillatory energy of the system (2.1) (defined with the original ω , not with $\tilde{\omega}$).

Theorem 8.1. *Let the Störmer–Verlet method be applied to the problem (2.1)–(2.3) with a step size h for which $0 < c_0 \leq h\omega \leq c_1 < 2$ and $|\sin(\frac{1}{2}kh\tilde{\omega})| \geq c\sqrt{h}$ for $k = 1, \dots, N$ for some $N \geq 2$ and $c > 0$. Suppose further that the numerical solution values x_n stay in a region on which all derivatives of U are bounded. Then, the modified energies along the numerical solution satisfy*

$$\begin{aligned} H^*(x_n, \dot{x}_n) &= H^*(x_0, \dot{x}_0) + \mathcal{O}(h) \\ I^*(x_n, \dot{x}_n) &= I^*(x_0, \dot{x}_0) + \mathcal{O}(h) \end{aligned} \quad \text{for } 0 \leq nh \leq h^{-N+1}. \quad (8.5)$$

The constants symbolized by \mathcal{O} are independent of n, h, ω with the above conditions.

Proof. With the modified velocities x'_n defined by

$$2h \operatorname{sinc}(h\tilde{\omega}) x'_n = x_{n+1} - x_{n-1}$$

method (8.2) becomes a method (2.6) with (2.11), or equivalently (2.7)-(2.8), with $\tilde{\omega}$ instead of ω and with $\psi(\xi) = \phi(\xi) = 1$.

The condition $0 < c_0 \leq h\omega \leq c_1 < 2$ implies $|\sin(\frac{1}{2}kh\tilde{\omega})| \geq c_2 > 0$ for $k = 1, 2$, and hence conditions (7.1) are trivially satisfied with $h\tilde{\omega}$ instead of $h\omega$. We are thus in the position to apply Theorem 7.1, which yields

$$\begin{aligned} \tilde{H}(x_n, x'_n) &= \tilde{H}(x_0, x'_0) + \mathcal{O}(h) \\ \tilde{I}(x_n, x'_n) &= \tilde{I}(x_0, x'_0) + \mathcal{O}(h) \end{aligned} \quad \text{for } 0 \leq nh \leq h^{-N+1}, \quad (8.6)$$

where \tilde{H} and \tilde{I} are defined in the same way as H and I , but with $\tilde{\omega}$ in place of ω . The components of the Störmer–Verlet velocities \dot{x}_n and the modified velocities x'_n are related by

$$\dot{x}_{n,0} = x'_{n,0}, \quad \dot{x}_{n,1} = \operatorname{sinc}(h\tilde{\omega}) x'_{n,1} = \frac{\omega}{\tilde{\omega}} \sqrt{1 - \frac{1}{4}h^2\omega^2} x'_{n,1}, \quad (8.7)$$

so that

$$\begin{aligned} \tilde{I}(x_n, x'_n) &= \frac{1}{2} \|x'_{n,1}\|^2 + \frac{1}{2} \tilde{\omega}^2 \|x_{n,1}\|^2 \\ &= \frac{1}{2} \frac{\tilde{\omega}^2}{\omega^2} \frac{1}{1 - \frac{1}{4}h^2\omega^2} \|\dot{x}_{n,1}\|^2 + \frac{1}{2} \frac{\tilde{\omega}^2}{\omega^2} \omega^2 \|x_{n,1}\|^2 \\ &= \frac{\tilde{\omega}^2}{\omega^2} I^*(x_n, \dot{x}_n). \end{aligned} \quad (8.8)$$

Similarly,

$$\begin{aligned} H^*(x_n, \dot{x}_n) &= \frac{1}{2} \|\dot{x}_{n,0}\|^2 + U(x_n) + I^*(x_n, \dot{x}_n) \\ &= \tilde{H}(x_n, x'_n) + \left(\frac{\omega^2}{\tilde{\omega}^2} - 1 \right) \tilde{I}(x_n, x'_n), \end{aligned} \quad (8.9)$$

and hence (8.6) yields the result. \square

For fixed $h\omega \geq c_0 > 0$ and $h \rightarrow 0$, the maximum deviation in the energy does not tend to 0, due to the highly oscillatory term $\frac{1}{2}\gamma\|\dot{x}_1\|^2$ in $H^*(x, \dot{x})$ and $I^*(x, \dot{x})$. We show, however, that *time averages* of H and I are nearly preserved over long time. For an arbitrary fixed $T > 0$, consider the averages over intervals of length T ,

$$\begin{aligned}
\overline{H}_n &= \frac{1}{T} h \sum_{|jh| \leq T/2} H(x_{n+j}, \dot{x}_{n+j}) \\
\overline{I}_n &= \frac{1}{T} h \sum_{|jh| \leq T/2} I(x_{n+j}, \dot{x}_{n+j}) .
\end{aligned} \tag{8.10}$$

Theorem 8.2. *Under the conditions of Theorem 8.1, the time averages of the total and the oscillatory energy along the numerical solution satisfy*

$$\begin{aligned}
\overline{H}_n &= \overline{H}_0 + \mathcal{O}(h) \\
\overline{I}_n &= \overline{I}_0 + \mathcal{O}(h) \quad \text{for } 0 \leq nh \leq h^{-N+1} .
\end{aligned} \tag{8.11}$$

The constants symbolized by \mathcal{O} are independent of n , h , ω with the above conditions.

Proof. We show

$$\begin{aligned}
\overline{H}_n &= H^*(x_n, \dot{x}_n) - \frac{1}{2} \frac{\gamma}{1+\gamma} I^*(x_n, \dot{x}_n) + \mathcal{O}(h) \\
\overline{I}_n &= I^*(x_n, \dot{x}_n) - \frac{1}{2} \frac{\gamma}{1+\gamma} I^*(x_n, \dot{x}_n) + \mathcal{O}(h) ,
\end{aligned} \tag{8.12}$$

which implies the result by Theorem 8.1. Consider the modulated Fourier expansions of x_n and x'_n for $t = nh$ in a bounded interval. Theorem 5.3 shows that

$$x'_{n,1} = i\tilde{\omega} (e^{i\tilde{\omega}t} z_{h,1}^1(t) - e^{-i\tilde{\omega}t} \overline{z_{h,1}^1(t)}) + \mathcal{O}(h) , \quad t = nh ,$$

with $z_{h,1}^1(t)$ from the modulated Fourier expansion of Theorem 5.2 (with $\tilde{\omega}$ instead of ω). With (8.7) it follows that

$$\dot{x}_{n,1} = i\omega \sqrt{1 - \frac{1}{4}h^2\omega^2} (e^{i\tilde{\omega}t} z_{h,1}^1(t) - e^{-i\tilde{\omega}t} \overline{z_{h,1}^1(t)}) + \mathcal{O}(h) ,$$

and therefore, recalling the definition of γ ,

$$\|\dot{x}_{n,1}\|^2 = \omega^2 \frac{1}{1+\gamma} \left(2 \|z_{h,1}^1(t)\|^2 - 2 \operatorname{Re} e^{2i\tilde{\omega}t} z_{h,1}^1(t)^2 \right) + \mathcal{O}(h) .$$

Theorems 5.2 and 5.3 yield

$$2\tilde{\omega}^2 \|z_{h,1}^1(t)\|^2 = \tilde{I}(x_n, x'_n) + \mathcal{O}(h)$$

and hence, by (8.8),

$$2\omega^2 \|z_{h,1}^1(t)\|^2 = I^*(x_n, \dot{x}_n) + \mathcal{O}(h) .$$

A partial summation shows that the time average over the highly oscillatory terms $e^{2i\tilde{\omega}t} \omega^2 z_{h,1}^1(t)^2$ is $\mathcal{O}(h)$. This finally gives

$$\frac{1}{T} h \sum_{|j| \leq T/2} \|\dot{x}_{j,1}\|^2 = \frac{1}{1+\gamma} I^*(x_n, \dot{x}_n) + \mathcal{O}(h) .$$

Taking the time averages in the expressions of the definition (8.4) of H^* and I^* then yields (8.12). \square

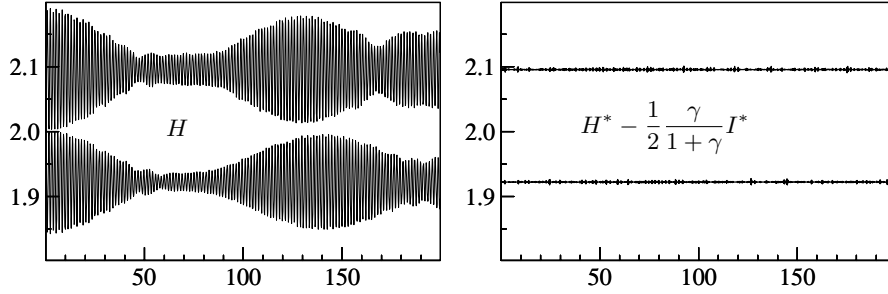


Fig. 8.1. Total energies (left) and their predicted averages (right) for the Störmer–Verlet method and for two different initial values, with $\omega = 50$ and h such that $h\omega = 0.8$

Figure 8.1 illustrates the above result. It shows the total energy H for two different initial values on the left, and the averages as predicted by the expression on the right-hand side of (8.12) on the right picture. The initial values are as in Chap. I with the exception of $x_{1,1}(0)$ and $\dot{x}_{1,1}(0)$. We take $x_{1,1}(0) = \sqrt{2}/\omega$, $\dot{x}_{1,1}(0) = 0$ for one set of initial values and $x_{1,1}(0) = 0$, $\dot{x}_{1,1}(0) = \sqrt{2}$ for the other. The total energies at the initial values are 2.00240032 and 2, respectively.

XIII.9 Systems with Several Constant Frequencies

This section studies the conservation of invariants and almost-invariants along numerical approximations of an extension of (2.1) to systems with the Hamiltonian function

$$H(p, q) = \frac{1}{2} p^T M^{-1} p + \frac{1}{2\varepsilon^2} q^T A q + U(q) \quad (9.1)$$

with a positive definite constant matrix M and a positive semi-definite constant matrix A . With the Cholesky decomposition $M = LL^T$ and the canonical transformation $\tilde{p} = L^{-1}p$, $\tilde{q} = L^T q$ we obtain a Hamiltonian where the mass matrix is the identity matrix and A is transformed to $\tilde{A} = L^{-1}AL^T$. Diagonalizing $\tilde{A} = Q\Lambda Q^T$ and transforming to $x = Q^T \tilde{q}$ then yields a Hamiltonian of the form (we omit the tilde on $\tilde{U}(x) = U(q)$ and $\tilde{H}(x, \dot{x}) = H(p, q)$)

$$H(x, \dot{x}) = \frac{1}{2} \sum_{j=0}^{\ell} \left(\|\dot{x}_j\|^2 + \frac{\lambda_j^2}{\varepsilon^2} \|x_j\|^2 \right) + U(x), \quad (9.2)$$

where $x = (x_0, x_1, \dots, x_\ell)$ with $x_j \in \mathbb{R}^{d_j}$, $\lambda_0 = 0$, and $\lambda_j > 0$ for $j \geq 1$ are all distinct. After rescaling ε we may assume $\lambda_j \geq 1$ for $j = 1, \dots, \ell$.

Following Cohen, Hairer & Lubich (2004) we extend the results of the previous sections to the multi-frequency case $\ell > 1$. Modulated Fourier expansions are again the basic analytical tool. A new aspect is possible resonance among the λ_j .

XIII.9.1 Oscillatory Energies and Resonances

The equations of motion for the Hamiltonian system (9.2) can be written as the system of second-order differential equations

$$\ddot{x} = -\Omega^2 x + g(x), \quad (9.3)$$

where $\Omega = \text{diag}(\omega_j I)$ with the frequencies $\omega_j = \lambda_j/\varepsilon$ and $g(x) = -\nabla U(x)$. As suitable numerical methods we consider again the class of trigonometric integrators studied in Sect. XIII.2, (2.6) with (2.11), with filter functions ψ and ϕ .

We are interested in the long-time near-conservation of the total energy $H(x, \dot{x})$ and the oscillatory energies

$$I_j(x, \dot{x}) = \frac{1}{2} \left(\|\dot{x}_j\|^2 + \frac{\lambda_j^2}{\varepsilon^2} \|x_j\|^2 \right) \quad \text{for } j \geq 1 \quad (9.4)$$

or suitable linear combinations thereof. Benettin, Galgani & Giorgilli (1989) have shown that the quantities

$$I_\mu(x, \dot{x}) = \sum_{j=1}^{\ell} \frac{\mu_j}{\lambda_j} I_j(x, \dot{x}) \quad (9.5)$$

are approximately preserved along every bounded solution of the Hamiltonian system that has a total energy bounded independently of ε , on exponentially long time intervals of size $\mathcal{O}(e^{c/\varepsilon})$ if the potential $U(x)$ is analytic and $\mu = (\mu_1, \dots, \mu_\ell)$ is orthogonal to the *resonance module*

$$\mathcal{M} = \{k \in \mathbb{Z}^\ell : k_1 \lambda_1 + \dots + k_\ell \lambda_\ell = 0\}, \quad (9.6)$$

if a diophantine non-resonance condition holds outside \mathcal{M} . (Cf. also Sect. XIII.9.4 below.)

Since $\mu = \lambda$ is orthogonal to \mathcal{M} , the total oscillatory energy $\sum_{j=1}^{\ell} I_j(x, \dot{x})$ of the system is approximately preserved independently of the resonance module \mathcal{M} . Subtracting this expression from the total energy (1.7), we see that also the *smooth energy*

$$K(x, \dot{x}) = \frac{1}{2} \|\dot{x}_0\|^2 + U(x) \quad (9.7)$$

is approximately preserved. With an ε -independent bound of the total energy $H(x, \dot{x})$ we have $x_j = \mathcal{O}(\varepsilon)$ for $j = 1, \dots, \ell$, so that $K(x, \dot{x})$ is close to the Hamiltonian of the reduced system in which all oscillatory degrees of freedom are taken out, $H_0(x_0, \dot{x}_0) = \frac{1}{2} \|\dot{x}_0\|^2 + U(x_0, 0, \dots, 0)$.

Example 9.1. To illustrate the conservation of the various energies, we consider a Hamiltonian (1.7) with $\ell = 3$, $\lambda = (1, \sqrt{2}, 2)$ and we assume that the dimensions of x_j are all 1 with the exception of that of $x_1 = (x_{1,1}, x_{1,2})$ which is 2. The resonance module is then $\mathcal{M} = \{(k_1, 0, k_3) : k_1 + 2k_3 = 0\}$. We take $\varepsilon^{-1} = \omega = 70$, the potential

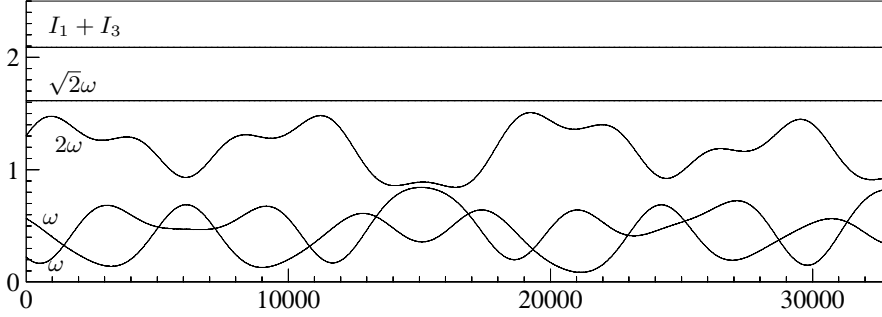


Fig. 9.1. Oscillatory energies of the individual components (the frequencies $\lambda_j\omega = \lambda_j/\varepsilon$ are indicated) and the sum $I_1 + I_3$ of the oscillatory energies corresponding to the resonant frequencies ω and 2ω

$$U(x) = (0.05 + x_{1,1} + x_{1,2} + x_2 + 2.5x_3)^4 + \frac{1}{8}x_0^2 x_{1,1}^2 + \frac{1}{2}x_0^2, \quad (9.8)$$

and $x(0) = (1, 0.3\varepsilon, 0.8\varepsilon, -1.1\varepsilon, 0.7\varepsilon)$, $\dot{x}(0) = (-0.2, 0.6, 0.7, -0.9, 0.8)$ as initial values. We consider I_μ for $\mu = (1, 0, 2)$ and $\mu = (0, \sqrt{2}, 0)$, which are both orthogonal to \mathcal{M} . In Fig. 9.1 we plot the oscillatory energies for the individual components of the system. The corresponding frequencies are attached to the curves. We also plot the sum $I_1 + I_3$ of the three oscillatory energies corresponding to the resonant frequencies $1/\varepsilon$ and $2/\varepsilon$. We see that $I_1 + I_3$ as well as I_2 (which are I_μ for the above two vectors $\mu \perp \mathcal{M}$) are well conserved over long times up to small oscillations of size $\mathcal{O}(\varepsilon)$. There is an energy exchange between the two components corresponding to the same frequency $1/\varepsilon$, and on a larger scale an energy exchange between I_1 and I_3 .

Numerical Experiment. As a first method we take (2.6) with $\phi(\xi) = 1$ and $\psi(\xi) = \text{sinc}(\xi)$, and we apply it with large step sizes so that $h\omega = h/\varepsilon$ takes the values 1, 2, 4, and 8. Figure 9.2 shows the various oscillatory energies which can be compared to the exact values in Fig. 9.1. For all step sizes, the oscillatory energy corresponding to the frequency $\sqrt{2}\omega$ and the sum $I_1 + I_3$ are well conserved on long time intervals. Oscillations in these expressions increase with h . The energy exchange between resonant frequencies is close to that of the exact solution. We have not plotted the total energy $H(x_n, \dot{x}_n)$ nor the smooth energy $K(x_n, \dot{x}_n)$ of (9.7). Both are well conserved over long times.

We repeat this experiment with the method where $\phi(\xi) = 1$ and $\psi(\xi) = \text{sinc}^2(\xi/2)$ (Fig. 9.3). Only the oscillatory energy corresponding to $\sqrt{2}\omega$ is approximately conserved over long times. Neither the expression $I_1 + I_3$ nor the total energy (not shown) are conserved. The smooth energy $K(x_n, \dot{x}_n)$ is, however, well conserved.

Figure 9.4 shows the corresponding result for the method with $\phi(\xi) = \text{sinc}(\xi)$ and $\psi(\xi) = \text{sinc}(\xi)\phi(\xi)$. The oscillatory energy for $\sqrt{2}\omega$ and also $I_1 + I_3$ are well conserved. However, the energy exchange between the resonant frequencies is not correctly reproduced.

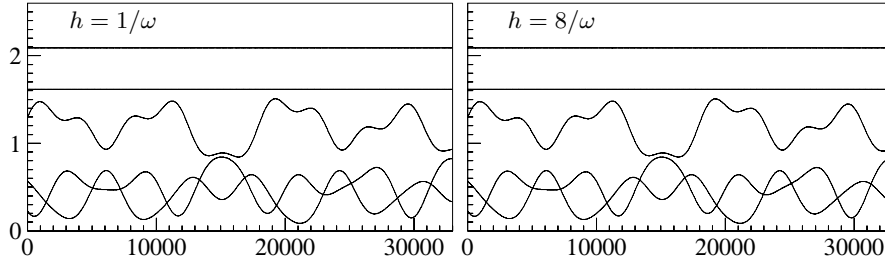


Fig. 9.2. Oscillatory energies as in Fig. 9.1 along the numerical solution of (2.6) with $\phi(\xi) = 1$ and $\psi(\xi) = \text{sinc}(\xi)$

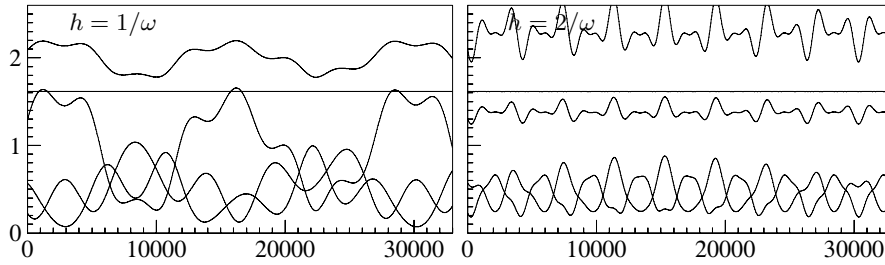


Fig. 9.3. Oscillatory energies as in Fig. 9.1 along the numerical solution of (2.6) with $\phi(\xi) = 1$ and $\psi(\xi) = \text{sinc}^2(\xi/2)$

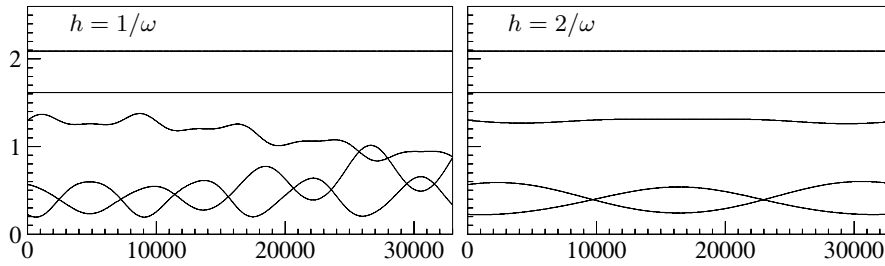


Fig. 9.4. Oscillatory energies as in Fig. 9.1 along the numerical solution of (2.6) with $\phi(\xi) = \text{sinc}(\xi)$ and $\psi(\xi) = \text{sinc}(\xi)\phi(\xi)$

XIII.9.2 Multi-Frequency Modulated Fourier Expansions

The above numerical phenomena can be understood with a multi-frequency version of the modulated Fourier expansions studied in the previous chapter. We only outline the derivation and properties, since they are in large parts similar to the single-frequency case. More details can be found in Cohen, Hairer & Lubich (2004). We assume conditions that extend those of the previous sections:

- The energy of the initial values is bounded independently of ε ,

$$\frac{1}{2}\|\dot{x}(0)\|^2 + \frac{1}{2}\|\Omega x(0)\|^2 \leq E. \quad (9.9)$$

- The numerical solution values Φx_n stay in a compact subset of a domain on which the potential U is smooth.
- We impose a lower bound on the step size: $h/\varepsilon \geq c_0 > 0$.
- We assume the numerical non-resonance condition

$$\left| \sin\left(\frac{h}{2\varepsilon} k \cdot \lambda\right) \right| \geq c \sqrt{h} \quad \text{for all } k \in \mathbb{Z}^\ell \setminus \mathcal{M} \text{ with } |k| \leq N, \quad (9.10)$$

for some $N \geq 2$ and $c > 0$.

- For the filter functions we assume that for $\xi_j = h\lambda_j/\varepsilon$ ($j = 1, \dots, \ell$),

$$\begin{aligned} |\psi(\xi_j)| &\leq C_1 \operatorname{sinc}^2(\tfrac{1}{2}\xi_j), \\ |\phi(\xi_j)| &\leq C_2 |\operatorname{sinc}(\tfrac{1}{2}\xi_j)|, \\ |\psi(\xi_j)| &\leq C_3 |\operatorname{sinc}(\xi_j)| |\phi(\xi_j)|. \end{aligned} \quad (9.11)$$

The conditions on the filter functions are somewhat stronger than necessary, but they facilitate the presentation in the following.

For a given vector $\lambda = (\lambda_1, \dots, \lambda_\ell)$ and for the resonance module \mathcal{M} defined by (9.6), we let \mathcal{K} be a set of representatives of the equivalence classes in $\mathbb{Z}^\ell/\mathcal{M}$ which are chosen such that for each $k \in \mathcal{K}$ the sum $|k| = |k_1| + \dots + |k_\ell|$ is minimal in the equivalence class $[k] = k + \mathcal{M}$, and with $k \in \mathcal{K}$, also $-k \in \mathcal{K}$. We denote, for N of (6.3),

$$\mathcal{N} = \{k \in \mathcal{K} : |k| < N\}, \quad \mathcal{N}^* = \mathcal{N} \setminus \{(0, \dots, 0)\}. \quad (9.12)$$

The following multi-frequency version of Theorem XIII.5.2 establishes a modulated Fourier expansion for the numerical solution.

Theorem 9.2. *Consider the numerical solution of the system (9.3) by the method (2.6) with step size h . Under conditions (9.9)–(9.11), the numerical solution admits an expansion*

$$x_n = y(t) + \sum_{k \in \mathcal{N}^*} e^{ik \cdot \omega t} z^k(t) + \Psi \cdot \mathcal{O}(t^2 h^N) \quad (9.13)$$

with $\omega = \lambda/\varepsilon$, uniformly for $0 \leq t = nh \leq T$ and ε and h satisfying $h/\varepsilon \geq c_0 > 0$. The modulation functions together with all their derivatives (up to some arbitrarily fixed order) are bounded by

$$\begin{aligned} y_0 &= \mathcal{O}(1), & y_j &= \mathcal{O}(\varepsilon^2) \\ z_j^{\pm \langle j \rangle} &= \mathcal{O}(\varepsilon), & \dot{z}_j^{\pm \langle j \rangle} &= \mathcal{O}(\varepsilon^2) \\ z_j^k &= \mathcal{O}(h\varepsilon^{|k|}) & \text{for } k \neq \pm \langle j \rangle \end{aligned} \quad (9.14)$$

for $j = 1, \dots, \ell$. Here, $\langle j \rangle = (0, \dots, 1, \dots, 0)$ is the j th unit vector. The last estimate holds also for z_0^k for all $k \in \mathcal{N}^*$. Moreover, the function y is real-valued and $z^{-k} = \overline{z^k}$ for all $k \in \mathcal{N}^*$. The constants symbolized by the \mathcal{O} -notation are independent of h , ε and λ_j with (9.10), but they depend on E , N , c , and T .

The proof extends that of Theorem XIII.5.2. In terms of the difference operator of the method, $L(hD) = e^{hD} - 2 \cos h\Omega + e^{-hD}$, the functions $y(t)$ and $z^k(t)$ are constructed such that, up to terms of size $\Psi \cdot \mathcal{O}(h^{N+2})$,

$$\begin{aligned} L(hD)y &= h^2 \Psi \left(g(\Phi y) + \sum_{s(\alpha) \sim 0} \frac{1}{m!} g^{(m)}(\Phi y)(\Phi z)^\alpha \right) \\ L(hD + i h k \cdot \omega) z^k &= h^2 \Psi \sum_{s(\alpha) \sim k} \frac{1}{m!} g^{(m)}(\Phi y)(\Phi z)^\alpha. \end{aligned}$$

Here, the sums on the right-hand side are over all $m \geq 1$ and over multi-indices $\alpha = (\alpha_1, \dots, \alpha_m)$ with $\alpha_j \in \mathcal{N}^*$, for which the sum $s(\alpha) = \sum_{j=1}^m \alpha_j$ satisfies the relation $s(\alpha) \sim k$, that is, $s(\alpha) - k \in \mathcal{M}$. The notation $(\Phi z)^\alpha$ is short for the m -tuple $(\Phi z^{\alpha_1}, \dots, \Phi z^{\alpha_m})$.

A similar expansion to that for x_n exists also for the velocity approximation \dot{x}_n , like in Theorem XIII.5.3. As a consequence, the oscillatory energy (9.4) along the numerical solution takes the form, at $t = nh \leq T$,

$$I_j(x_n, \dot{x}_n) = 2\omega_j^2 \|z_j^{(j)}(t)\|^2 + \mathcal{O}(\varepsilon). \quad (9.15)$$

With the first terms of the modulated Fourier expansion one proves, as in Theorems XIII.4.1 and XIII.4.2, error bounds over bounded time intervals which are of second order in the positions and of first order in the velocities:

$$\|x_n - x(t_n)\| \leq C h^2, \quad \|\dot{x}_n - \dot{x}(t_n)\| \leq C h, \quad (9.16)$$

where C is independent of ε , h and n with $nh \leq T$ and of bounds of solution derivatives.

XIII.9.3 Almost-Invariants of the Modulation System

With $y^0(t) = z^0(t) = y(t)$ and $y^k(t) = e^{ik \cdot \omega t} z^k(t)$ for $k \in \mathcal{N}$, where y and z^k are the modulation functions of Theorem 9.2, we denote

$$\mathbf{y} = (y^k)_{k \in \mathcal{N}}, \quad \mathbf{z} = (z^k)_{k \in \mathcal{N}}.$$

We introduce the extended potential

$$\mathcal{U}(\mathbf{y}) = U(\Phi y^0) + \sum_{s(\alpha) \sim 0} \frac{1}{m!} U^{(m)}(\Phi y^0)(\Phi \mathbf{y})^\alpha, \quad (9.17)$$

where the sum is again taken over all $m \geq 1$ and all multi-indices $\alpha = (\alpha_1, \dots, \alpha_m)$ with $\alpha_j \in \mathcal{N}^*$ for which $s(\alpha) = \sum_j \alpha_j \in \mathcal{M}$. The functions $y^k(t)$ then satisfy

$$\Psi^{-1} \Phi h^{-2} L(hD) y^k = -\nabla_{-k} \mathcal{U}(\mathbf{y}) + \Phi \cdot \mathcal{O}(h^N), \quad (9.18)$$

where ∇_{-k} denotes the gradient with respect to the variable y^{-k} . This system has almost-invariants that are related to the Hamiltonian H and the oscillatory energies I_μ with $\mu \perp \mathcal{M}$.

The Energy-Type Almost-Invariant of the Modulation System. We multiply (9.18) by $(\dot{y}^{-k})^T$ and sum over $k \in \mathcal{N}$ to obtain

$$\sum_{k \in \mathcal{N}} (\dot{y}^{-k})^T \Psi^{-1} \Phi h^{-2} L(hD) y^k + \frac{d}{dt} \mathcal{U}(\mathbf{y}) = \mathcal{O}(h^N).$$

Since we know bounds of the modulation functions z^k and of their derivatives from Theorem 9.2, we rewrite this relation in terms of the quantities z^k :

$$\sum_{k \in \mathcal{N}} (\dot{z}^{-k} - ik \cdot \omega z^{-k})^T \Psi^{-1} \Phi h^{-2} L(hD + i h k \cdot \omega) z^k + \frac{d}{dt} \mathcal{U}(\mathbf{z}) = \mathcal{O}(h^N). \quad (9.19)$$

As in (6.21) we obtain that the left-hand side of (9.19) can be written as the time derivative of a function $\mathcal{H}^*[\mathbf{z}](t)$ which depends on the values at t of the modulation-function vector \mathbf{z} and its first N time derivatives. The relation (9.19) thus becomes

$$\frac{d}{dt} \mathcal{H}^*[\mathbf{z}](t) = \mathcal{O}(h^N).$$

Together with the estimates of Theorem 9.2 this construction of \mathcal{H}^* yields the following multi-frequency extension of Lemma XIII.6.4.

Lemma 9.3. *Under the assumptions of Theorem 9.2, the modulation functions $\mathbf{z} = (z^k)_{k \in \mathcal{N}}$ of the numerical solution satisfy*

$$\mathcal{H}^*[\mathbf{z}](t) = \mathcal{H}^*[\mathbf{z}](0) + \mathcal{O}(th^N) \quad (9.20)$$

for $0 \leq t \leq T$. Moreover, at $t = nh$ we have

$$\mathcal{H}^*[\mathbf{z}](t) = H^*(x_n, \dot{x}_n) + \mathcal{O}(h), \quad (9.21)$$

where, with $\sigma(\xi) = \text{sinc}(\xi)\phi(\xi)/\psi(\xi)$ and $\xi_j = h\lambda_j/\varepsilon$,

$$H^*(x, \dot{x}) = H(x, \dot{x}) + \sum_{j=1}^{\ell} (\sigma(\xi_j) - 1) I_j(x, \dot{x}). \quad (9.22)$$

The Momentum-Type Almost-Invariants of the Modulation System. The equations (9.18) have further almost-invariants that result from invariance properties of the extended potential \mathcal{U} , similarly as the conservation of angular momentum results from an invariance of the potential U in a mechanical system by Noether's theorem. For $\mu \in \mathbb{R}^\ell$ and $\mathbf{y} = (y^k)_{k \in \mathcal{N}}$ we set

$$S_\mu(\tau)\mathbf{y} = (e^{ik \cdot \mu \tau} y^k)_{k \in \mathcal{N}}, \quad \tau \in \mathbb{R}$$

so that, by the multi-linearity of the derivative, the definition (9.17) yields

$$\mathcal{U}(S_\mu(\tau)\mathbf{y}) = U(\Phi y^0) + \sum_{s(\alpha) \sim 0} \frac{e^{is(\alpha) \cdot \mu \tau}}{m!} U^{(m)}(\Phi y^0)(\Phi \mathbf{y})^\alpha. \quad (9.23)$$

If $\mu \perp \mathcal{M}$, then the relation $s(\alpha) \sim 0$ implies $s(\alpha) \cdot \mu = 0$, and hence the expression (9.23) is independent of τ . It therefore follows that

$$0 = \frac{d}{d\tau} \Big|_{\tau=0} \mathcal{U}(S_\mu(\tau)\mathbf{y}) = \sum_{k \in \mathcal{N}} i(k \cdot \mu) (y^k)^T \nabla_k \mathcal{U}(\mathbf{y})$$

for all vectors $\mathbf{y} = (y^k)_{k \in \mathcal{N}}$. If μ is not orthogonal to \mathcal{M} , some terms in the sum of (9.23) depend on τ . However, for these terms with $s(\alpha) \in \mathcal{M}$ and $s(\alpha) \cdot \mu \neq 0$ we have $|s(\alpha)| \geq M = \min\{|k| : 0 \neq k \in \mathcal{M}\}$ and if $\mu \perp \mathcal{M}_N$, then $|s(\alpha)| \geq N+1$. The bounds (5.13) then yield

$$\sum_{k \in \mathcal{N}} i(k \cdot \mu) (y^k)^T \nabla_k \mathcal{U}(\mathbf{y}) = \begin{cases} \mathcal{O}(\varepsilon^M) & \text{for arbitrary } \mu \\ \mathcal{O}(\varepsilon^{N+1}) & \text{for } \mu \perp \mathcal{M}_N \end{cases} \quad (9.24)$$

for the vector $\mathbf{y} = \mathbf{y}(t)$ as given by Theorem 9.2. Multiplying the relation (9.18) by $\frac{i}{\varepsilon}(-k \cdot \mu) (y^{-k})^T$ and summing over $k \in \mathcal{N}$, we obtain with (9.24) that

$$-\frac{i}{\varepsilon} \sum_{k \in \mathcal{N}} (k \cdot \mu) (y^{-k})^T \Psi^{-1} \Phi h^{-2} L(hD) y^k = \mathcal{O}(h^N) + \mathcal{O}(\varepsilon^{M-1}).$$

The $\mathcal{O}(\varepsilon^{M-1})$ term is not present for $\mu \perp \mathcal{M}_N$. Written in the z variables, this becomes

$$-\frac{i}{\varepsilon} \sum_{k \in \mathcal{N}} (k \cdot \mu) (z^{-k})^T \Psi^{-1} \Phi h^{-2} L(hD + i h k \cdot \omega) z^k = \mathcal{O}(h^N) + \mathcal{O}(\varepsilon^{M-1}). \quad (9.25)$$

As in (9.19), the left-hand expression turns out to be the time derivative of a function $\mathcal{I}_\mu^*[\mathbf{z}](t)$ which depends on the values at t of the function \mathbf{z} and its first N derivatives:

$$\frac{d}{dt} \mathcal{I}_\mu^*[\mathbf{z}](t) = \mathcal{O}(h^N) + \mathcal{O}(\varepsilon^{M-1}).$$

Together with Theorem 9.2 this yields the following.

Lemma 9.4. *Under the assumptions of Theorem 9.2, the modulation functions \mathbf{z} satisfy*

$$\mathcal{I}_\mu^*[\mathbf{z}](t) = \mathcal{I}_\mu^*[\mathbf{z}](0) + \mathcal{O}(th^N) + \mathcal{O}(t\varepsilon^{M-1}) \quad (9.26)$$

for all $\mu \in \mathbb{R}^\ell$ and for $0 \leq t \leq T$. They satisfy

$$\mathcal{I}_\mu^*[\mathbf{z}](t) = \mathcal{I}_\mu^*[\mathbf{z}](0) + \mathcal{O}(th^N) \quad (9.27)$$

for $\mu \perp \mathcal{M}_N$ and $0 \leq t \leq T$. Moreover, at $t = nh$,

$$\mathcal{I}_\mu^*[\mathbf{z}](t) = I_\mu^*(x_n, \dot{x}_n) + \mathcal{O}(\varepsilon), \quad (9.28)$$

where, again with $\sigma(\xi) = \text{sinc}(\xi)\phi(\xi)/\psi(\xi)$,

$$I_\mu^*(x, \dot{x}) = \sum_{j=1}^{\ell} \sigma(\xi_j) \frac{\mu_j}{\lambda_j} I_j(x, \dot{x}). \quad (9.29)$$

XIII.9.4 Long-Time Near-Conservation of Total and Oscillatory Energies

With the proof of Theorem XIII.7.1, the above two lemmas yield the following results from Cohen, Hairer & Lubich (2004).

Theorem 9.5. *Under conditions (9.9)–(9.11), the numerical solution obtained by method (2.6) with (2.11) satisfies, for H^* and I_μ^* defined by (9.22) and (9.29),*

$$\begin{aligned} H^*(x_n, \dot{x}_n) &= H^*(x_0, \dot{x}_0) + \mathcal{O}(h) \\ I_\mu^*(x_n, \dot{x}_n) &= I_\mu^*(x_0, \dot{x}_0) + \mathcal{O}(h) \end{aligned} \quad \text{for } 0 \leq nh \leq h^{-N+1}$$

for $\mu \in \mathbb{R}^\ell$ with $\mu \perp \mathcal{M}_N = \{k \in \mathcal{M} : |k| \leq N\}$. The constants symbolized by \mathcal{O} are independent of $n, h, \varepsilon, \lambda_j$ satisfying the above conditions, but depend on N and the constants in the conditions.

Since $\mu = \lambda$ is always orthogonal to \mathcal{M} and to \mathcal{M}_N , the relation

$$K(x, \dot{x}) = H^*(x, \dot{x}) - I_\lambda^*(x, \dot{x})$$

for the smooth energy (9.7) implies

$$K(x_n, \dot{x}_n) = K(x_0, \dot{x}_0) + \mathcal{O}(h) \quad \text{for } 0 \leq nh \leq h^{-N+1}. \quad (9.30)$$

For $\sigma(\xi) = 1$ (or equivalently $\psi(\xi) = \text{sinc}(\xi)\phi(\xi)$) the modified energies H^* and I_μ^* are identical to the original energies H and I_μ of (9.2) and (9.5). The condition $\psi(\xi) = \text{sinc}(\xi)\phi(\xi)$ is known to be equivalent to the symplecticity of the one-step method $(x_n, \dot{x}_n) \mapsto (x_{n+1}, \dot{x}_{n+1})$, but its appearance in the above theorem is caused by a different mechanism which is not in any obvious way related to symplecticity. Without this condition we still have the following result, which also considers the long-time near-conservation of the individual oscillatory energies I_j for $j = 1, \dots, \ell$.

Theorem 9.6. *Under conditions (9.9)–(9.11), the numerical solution obtained by method (2.6) with (2.11) satisfies*

$$\begin{aligned} H(x_n, \dot{x}_n) &= H(x_0, \dot{x}_0) + \mathcal{O}(h) \\ I_j(x_n, \dot{x}_n) &= I_j(x_0, \dot{x}_0) + \mathcal{O}(h) \end{aligned} \quad \text{for } 0 \leq nh \leq h \cdot \min(\varepsilon^{-M+1}, h^{-N})$$

for $j = 1, \dots, \ell$, with $M = \min\{|k| : 0 \neq k \in \mathcal{M}\}$. The constants symbolized by \mathcal{O} are independent of $n, h, \varepsilon, \lambda_j$ satisfying the above conditions, but depend on N and the constants in the conditions.

For the non-resonant case $\mathcal{M} = \{0\}$ we have $M = \infty$ and hence the length of the interval with energy conservation is only restricted by (9.10). Notice that always $M \geq 3$, and $M = 3$ only in the case of a 1:2 resonance among the λ_j . For a 1:3 resonance we have $M = 4$ and in all other cases $M \geq 5$.

Explanation of the Numerical Experiment of Sect. XIII.9.1. All numerical methods in Figs. 9.2–9.4 satisfy the conditions of Theorems 9.6 and 9.5 for the step sizes considered.

In Fig. 9.2 we have the (symplectic) method (2.6) with $\phi(\xi) = 1$ and $\psi(\xi) = \text{sinc}(\xi)$, which has $\sigma(\xi) = 1$, so that H and H^* , and I_μ and I_μ^* coincide. For all step sizes, the oscillatory energy I_2 corresponding to the non-resonant frequency $\sqrt{2}\omega$ and the sum $I_1 + I_3$ are well conserved on long time intervals, in accordance with Theorem 9.5. The individual energies I_1 and I_3 corresponding to the resonant frequencies $\omega = 1/\varepsilon$ and $2/\varepsilon$ are not preserved on the time scale considered here, cf. Fig. 9.1. In fact, Theorem 9.6 here yields only a time scale $\mathcal{O}(h\varepsilon^{-2})$.

In Fig. 9.3 we use the method with $\phi(\xi) = 1$ and $\psi(\xi) = \text{sinc}^2(\xi/2)$, for which $\sigma(\xi)$ is not identical to 1, and hence H and H^* , and I_μ and I_μ^* do not coincide. The oscillatory energy $I_2 = \sigma_2^{-1} I_\mu^*$ with $\mu = (0, 1, 0) \perp \mathcal{M}$, which corresponds to the non-resonant frequency $\sqrt{2}\omega$, is approximately conserved over long times. Since Theorem 9.5 only states that the *modified* energies are well preserved, it is not surprising that neither $I_1 + I_3$ nor the original total energy H (not shown in the figure) are conserved. The modified energies H^* and $\sigma_1 I_1 + \sigma_3 I_3$ (not shown) are indeed well conserved, and so is the smooth energy K , in agreement with (9.30).

Figure 9.4 shows the result for the (symplectic) method with $\phi(\xi) = \text{sinc}(\xi)$ and $\psi(\xi) = \text{sinc}(\xi)\phi(\xi)$. Since $\sigma(\xi) = 1$, the oscillatory energy I_2 for $\sqrt{2}\omega$ and also $I_1 + I_3$ are well conserved, in agreement with Theorem 9.5. However, the energy exchange between the resonant frequencies is not correctly reproduced. This behaviour is not explained by Theorems 9.5 and 9.6, but it corresponds to the analysis in Sect. XIII.4.2 which, for the single-frequency case, explains the incorrect energy exchange of methods that do not satisfy $\psi(\xi)\phi(\xi) = \text{sinc}(\xi)$ (and thus, of all symplectic methods (2.7)–(2.10), with the exception of the above method with $\phi(\xi) = 1$ and $\psi(\xi) = \text{sinc}(\xi)$). That analysis could be extended to the multi-frequency case considered here.

We remark that the techniques of Sects. XIII.9.2 and XIII.9.3 can also be used to study the energy error of the Störmer–Verlet method, as in Sect. XIII.8; see Theorem 5.1 in Cohen, Hairer & Lubich (2004). The modulated Fourier expansion of the exact solution yields results on the near-preservation of the oscillatory energies along a bounded exact solution: under the energy bound (9.9) and the non-resonance condition

$$|k \cdot \lambda| \geq c\sqrt{\varepsilon} \quad \text{for } k \in \mathbb{Z}^\ell \setminus \mathcal{M} \text{ with } |k| \leq N \quad (9.31)$$

we have (see Theorem 6.1 in Cohen, Hairer & Lubich 2004)

$$I_\mu(x(t), \dot{x}(t)) = I_\mu(x(0), \dot{x}(0)) + \mathcal{O}(\varepsilon) \quad \text{for } 0 \leq t \leq \varepsilon^{-N+1} \quad (9.32)$$

for $\mu \in \mathbb{R}^\ell$ with $\mu \perp \mathcal{M}_N = \{k \in \mathcal{M} : |k| \leq N\}$. We further have

$$I_j(x(t), \dot{x}(t)) = I_j(x(0), \dot{x}(0)) + \mathcal{O}(\varepsilon) \quad \text{for } 0 \leq t \leq \varepsilon \cdot \min(\varepsilon^{-M+1}, \varepsilon^{-N}) \quad (9.33)$$

for $j = 1, \dots, \ell$, with $M = \min\{|k| : 0 \neq k \in \mathcal{M}\}$.

XIII.10 Systems with Non-Constant Mass Matrix

The high frequencies of the linearized differential equation remain constant up to small deviations for mechanical systems with a Hamiltonian of the form

$$H(p, q) = \frac{1}{2} p_0^T M_0(q)^{-1} p_0 + \frac{1}{2} p_1^T M_1^{-1} p_1 + \frac{1}{2} p^T R(q) p + \frac{1}{2\varepsilon^2} q_1^T A_1 q_1 + U(q) \quad (10.1)$$

with a symmetric positive definite matrix $M_0(q)$, constant symmetric positive definite matrices M_1 and A_1 , a symmetric matrix $R(q)$ with

$$R(q_0, 0) = 0,$$

and a potential $U(q)$. All the functions are assumed to depend smoothly on q . Bounded energy then requires $q_1 = \mathcal{O}(\varepsilon)$, so that $p^T R(q) p = \mathcal{O}(\varepsilon)$, but the derivative of this term with respect to q_1 is $\mathcal{O}(1)$.

As in (9.1), we may assume, after an appropriate canonical linear transformation based on a Cholesky decomposition of the mass matrix and a diagonalization of the resulting stiffness matrix, that the Hamiltonian is of the form

$$H(p, q) = \frac{1}{2} p_0^T M_0(q)^{-1} p_0 + \frac{1}{2} \sum_{j=1}^{\ell} \left(\|p_j\|^2 + \frac{\lambda_j^2}{\varepsilon^2} \|q_j\|^2 \right) + \frac{1}{2} p^T R(q) p + U(q) \quad (10.2)$$

with distinct, constant $\lambda_j \geq 1$.

The necessity for such a generalization results from the fact that oscillatory mechanical systems with near-constant frequencies in 2 or 3 space dimensions typically cannot be put in the form (9.1), but in the more general form (10.1) or (10.2).

Example 10.1 (Stiff Spring Pendulum). The motion of a mass point (of mass 1) hanging on a massless stiff spring (with spring constant $1/\varepsilon^2$) is described in polar coordinates $x_1 = r \sin \varphi$, $x_2 = -r \cos \varphi$ by the Lagrangian with kinetic energy $T = \frac{1}{2} (\dot{x}_1^2 + \dot{x}_2^2) = \frac{1}{2} (\dot{r}^2 + r^2 \dot{\varphi}^2)$ and potential energy $U = \frac{1}{2\varepsilon^2} (r-1)^2 - r \cos \varphi$. With the coordinates $q_0 = \varphi$, $q_1 = r-1$ and the conjugate momenta $p_i = \partial T / \partial \dot{q}_i$ this gives the Hamiltonian

$$H(p, q) = \frac{1}{2} ((1+q_1)^{-2} p_0^2 + p_1^2) + \frac{1}{2\varepsilon^2} q_1^2 - (1+q_1) \cos q_0,$$

which is of the form (10.2).

Numerical methods for systems (10.2) are studied by Cohen (2005). He splits the small term $\frac{1}{2}p^T R(q)p$ from the principal terms of the Hamiltonian and proposes the following method, where

$$K(p_0, q) = \frac{1}{2}p_0^T M_0(q)^{-1}p_0 + U(q).$$

Algorithm 10.2. 1. A half-step with the symplectic Euler method applied to the system with Hamiltonian $\frac{1}{2}p^T R(q)p$ gives

$$\begin{aligned}\hat{p}^n &= p^n - \frac{h}{2} \nabla_q \left(\frac{1}{2} (\hat{p}^n)^T R(q^n) \hat{p}^n \right) \\ \hat{q}^n &= q^n + \frac{h}{2} R(q^n) \hat{p}^n.\end{aligned}\tag{10.3}$$

2. Treating the oscillatory components of the variables p and q with a trigonometric method (2.7)–(2.8) and the slow components with the Störmer-Verlet scheme yields (for $j = 1, \dots, \ell$ and with $\omega_j = \lambda_j/\varepsilon$ and $\xi_j = h\omega_j$)

$$\begin{aligned}p_0^{n+1/2} &= \hat{p}_0^n - \frac{h}{2} \nabla_{q_0} K(p_0^{n+1/2}, \Phi \hat{q}^n) \\ \hat{q}_0^{n+1} &= \hat{q}_0^n + \frac{h}{2} \left(\nabla_{p_0} K(p_0^{n+1/2}, \Phi \hat{q}^n) + \nabla_{p_0} K(p_0^{n+1/2}, \Phi \hat{q}^{n+1}) \right) \\ \hat{q}_j^{n+1} &= \cos(\xi_j) \hat{q}_j^n + \omega_j^{-1} \sin(\xi_j) \hat{p}_j^n - \frac{h^2}{2} \psi(\xi_j) \nabla_{q_j} K(p_0^{n+1/2}, \Phi \hat{q}^n) \\ \hat{p}_j^{n+1} &= -\omega_j \sin(\xi_j) \hat{q}_j^n + \cos(\xi_j) \hat{p}_j^n - \frac{h}{2} \left(\psi_0(\xi_j) \nabla_{q_j} K(p_0^{n+1/2}, \Phi \hat{q}^n) \right. \\ &\quad \left. + \psi_1(\xi_j) \nabla_{q_j} K(p_0^{n+1/2}, \Phi \hat{q}^{n+1}) \right), \\ \hat{p}_0^{n+1} &= p_0^{n+1/2} - \frac{h}{2} \nabla_{q_0} K(p_0^{n+1/2}, \Phi \hat{q}^{n+1})\end{aligned}\tag{10.4}$$

where $\Phi = \phi(h\Omega)$ with $\Omega = \text{diag}(\omega_j I)$.

3. A half-step with the adjoint symplectic Euler method applied to the system with Hamiltonian $\frac{1}{2}p^T R(q)p$ gives

$$\begin{aligned}p^{n+1} &= \hat{p}^{n+1} - \frac{h}{2} \nabla_q \left(\frac{1}{2} (\hat{p}^{n+1})^T R(q^{n+1}) \hat{p}^{n+1} \right) \\ q^{n+1} &= \hat{q}^{n+1} + \frac{h}{2} R(q^{n+1}) \hat{p}^{n+1}.\end{aligned}\tag{10.5}$$

The filter functions $\psi, \psi_0, \psi_1, \phi$ are again real-valued functions with $\psi(0) = \hat{\psi}(0) = \tilde{\psi}(0) = \phi(0) = 1$ that satisfy (2.9). The method is still symplectic if and only if (2.10) holds. Note that Step 2. of the algorithm is explicit if $M_0(q)$ does not depend on q_0 .

Cohen (2004, 2005) studies the modulated Fourier expansion of this method and shows that the long-time near-conservation of total and oscillatory energies as given by Theorem 9.6 remains valid also in this more general situation.

Example 10.3 (Triatomic Molecule). The motion of a near-rigid triatomic molecule is described by a Hamiltonian system with a Hamiltonian (10.2). For simplicity we fix the position of the central atom. We then have two stiff-spring pendulums strongly coupled by another spring. With angles and distances as shown in Fig. 10.1, we use the position coordinates $\varphi_1, q_1 = r_1 - 1, \varphi_2, q_2 = r_2 - 1$ with the conjugate momenta π_1, p_1, π_2, p_2 , respectively. The Hamiltonian then reads

$$H(\pi, p, \varphi, q) = \frac{1}{2} \left((1 + q_1)^{-2} \pi_1^2 + p_1^2 + (1 + q_2)^{-2} \pi_2^2 + p_2^2 \right) + \frac{1}{2\varepsilon^2} \left(q_1^2 + q_2^2 + \frac{\alpha^2}{2} (\varphi_2 - \varphi_1)^2 \right) + U(\varphi, q) \quad (10.6)$$

with a spring constant $\frac{1}{2}\alpha^2/\varepsilon^2$ for connecting the two pendulums and an external potential U . With the canonical change of variables

$$\begin{pmatrix} q_3 \\ q_0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}, \quad \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} p_3 \\ p_0 \end{pmatrix},$$

the Hamiltonian takes the form (10.2):

$$H(p, q) = \frac{1}{2} (p_0^2 + p_1^2 + p_2^2 + p_3^2) + \frac{1}{2\varepsilon^2} (q_1^2 + q_2^2 + \alpha^2 q_3^2) + p^T R(q) p + \hat{U}(q) \quad (10.7)$$

with

$$p^T R(q) p = -\frac{1}{4} \frac{2q_2 + q_2^2}{(1 + q_2)^2} (p_0 - p_3)^2 - \frac{1}{4} \frac{2q_1 + q_1^2}{(1 + q_1)^2} (p_0 + p_3)^2$$

and $\hat{U}(q) = U(\varphi_1, \varphi_2, q_1, q_2)$.

For the water molecule the ratio between the frequencies of the bond angle and the bond lengths is $\alpha \approx 0.2$, according to some popular models. In our numerical experiments, we observed good conservation of all the oscillatory energies and the total energy. More interesting phenomena occur in a near-resonance situation. We consider $\alpha = 0.49$ and $\varepsilon = 0.01$, no exterior potential ($U = 0$), and initial values $q(0) = (0, \varepsilon/2, \varepsilon, \alpha/\varepsilon)$ and $p(0) = (1.1, 0.2, -0.8, 1.3)$. In Fig. 10.2 we apply the method of Algorithm 10.2 with step sizes $h = 0.5\varepsilon$ and $h = 2\varepsilon$ and obtain

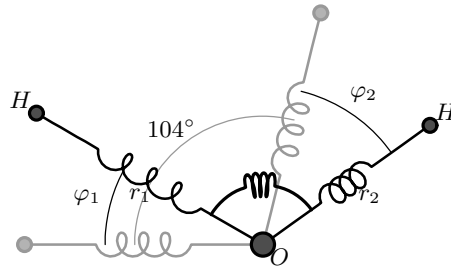


Fig. 10.1. Water molecule and reference configuration as gray shadow

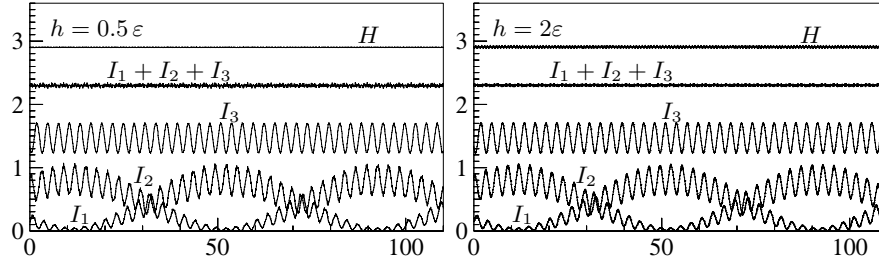


Fig. 10.2. Oscillatory energies and total energy for the method of Algorithm 10.2

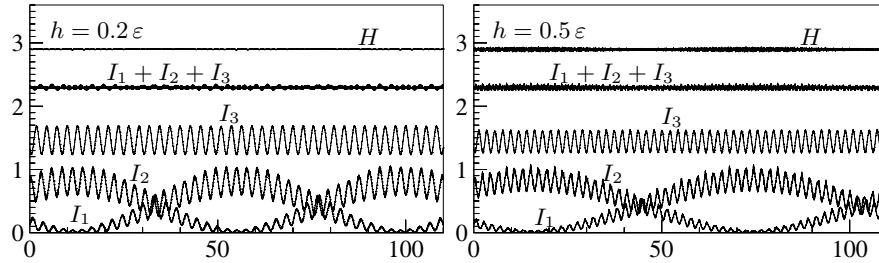


Fig. 10.3. Oscillatory energies and total energy for the Störmer–Verlet method

numerical results that agree very well with a solution obtained with very small step sizes. For comparison we show in Fig. 10.3 the results of the Störmer–Verlet method with step sizes $h = 0.2\varepsilon$ and $h = 0.5\varepsilon$, for which the energy exchange is not correct. For the reason explained in Sect. VI.3, (3.2)–(3.3), both methods are fully explicit for this problem.

XIII.11 Exercises

1. Show that the impulse method (with exact solution of the fast system) reduces to Deuffhard's method in the case of a quadratic potential $W(q) = \frac{1}{2}q^T Aq$.
2. Show that a method (2.7)–(2.8) satisfying (2.9) is symplectic if and only if

$$\psi(\xi) = \text{sinc}(\xi) \phi(\xi) \quad \text{for } \xi = h\omega.$$

3. The change of coordinates $x_n = \chi(h\Omega)z_n$ transforms (2.7)–(2.8) into a method of identical form with $\phi, \psi, \psi_0, \psi_1$ replaced by $\chi\phi, \chi^{-1}\psi, \chi^{-1}\psi_0, \chi^{-1}\psi_1$. Prove that, for $h\omega$ satisfying $\text{sinc}(h\omega)\phi(h\omega)/\psi(h\omega) > 0$, it is possible to find $\chi(h\omega)$ such that the transformed method is symplectic.
4. Prove that for infinitely differentiable functions $g(t)$ the solution of $\ddot{x} + \omega^2 x = g(t)$ can be written as

$$x(t) = y(t) + \cos(\omega t) u(t) + \sin(\omega t) v(t),$$

where $y(t)$, $u(t)$, $v(t)$ are given by asymptotic expansions in powers of ω^{-1} .

Hint. Use the variation-of-constants formula and apply repeated partial integration.

5. Show that the recurrence relation $e_{n+1} - 2\cos(h\Omega)e_n + e_{n-1} = b_n$ has the solution

$$e_{n+1} = -W_{n-1}e_0 + W_n e_1 + \sum_{j=1}^n W_{n-j} b_j$$

with $W_n = \sin(h\Omega)^{-1} \sin((n+1)h\Omega)$ (or the appropriate limit when $\sin(h\Omega)$ is not invertible).

6. Consider a Hamiltonian $H(p_R, p_I, q_R, q_I)$ and let

$$\mathcal{H}(p, q) = 2H(p_R, p_I, q_R, q_I)$$

for $p = p_R + ip_I$, $q = q_R + iq_I$. Prove that in the new variables p, q the Hamiltonian system becomes

$$\dot{p} = -\frac{\partial \mathcal{H}}{\partial \bar{q}}(p, q), \quad \dot{q} = \frac{\partial \mathcal{H}}{\partial p}(p, q).$$

7. Prove the following refinement of Theorem 6.3: along the solution $x(t)$ of (2.1), the modified oscillatory energy $J(x, \dot{x}) = I(x, \dot{x}) - x_1^T g_1(x)$ satisfies

$$J(x(t), \dot{x}(t)) = J(x(0), \dot{x}(0)) + \mathcal{O}(\omega^{-2}) + \mathcal{O}(t\omega^{-N}).$$

8. Define $\hat{H}(x, \dot{x}) = H(x, \dot{x}) - \rho x_1^T g_1(x)$, $\hat{J}(x, \dot{x}) = J(x, \dot{x}) - \rho x_1^T g_1(x)$ with $J(x, \dot{x})$ of the previous exercise and with

$$\rho = \frac{\psi(h\omega)}{\text{sinc}^2(\frac{1}{2}h\omega)} - 1.$$

In the situation of Theorem 7.1, show that

$$\begin{aligned} \hat{H}(x_n, \dot{x}_n) &= \hat{H}(x_0, \dot{x}_0) + \mathcal{O}(h^2) \\ \hat{J}(x_n, \dot{x}_n) &= \hat{J}(x_0, \dot{x}_0) + \mathcal{O}(h^2) \end{aligned} \quad \text{for } 0 \leq nh \leq h^{-N+1}.$$

Notice that the total energy $H(x_n, \dot{x}_n)$ and the modified oscillatory energy $J(x_n, \dot{x}_n)$ are conserved up to $\mathcal{O}(h^2)$ if $\rho = 0$, i.e., if $\psi(\xi) = \text{sinc}^2(\frac{1}{2}\xi)$. This explains the excellent energy conservation of methods (A) and (D) in Figure 2.5 away from resonances.

9. Generalizing the analysis of Sect. XIII.8, study the energy behaviour of the impulse or averaged-force multiple time-stepping method of Sect. VIII.4 with a fixed number N of Störmer–Verlet substeps per step, when the method is applied to the model problem with $h\omega$ bounded away from zero.

Chapter XIV.

Oscillatory Differential Equations with Varying High Frequencies

New aspects come into play when the high frequencies in an oscillatory system and their associated eigenspaces do not remain nearly constant, as in the previous chapter, but change with time or depend on the solution. We begin by studying linear differential equations with a time-dependent skew-hermitian matrix and then turn to nonlinear oscillatory mechanical systems with time- or solution-dependent frequencies. Our analysis uses canonical coordinate transforms that separate slow and fast motions and relate the fast oscillations to the skew-hermitian linear case. For the numerical treatment we consider suitably constructed long-time-step methods (“adiabatic integrators”) and multiple time-stepping methods.

XIV.1 Linear Systems with Time-Dependent Skew-Hermitian Matrix

We consider first-order linear differential equations with a skew-hermitian matrix that changes slowly compared to the rapid oscillations in the solution, a problem that has attracted much attention in quantum mechanics. We present a suitable class of numerical methods, termed adiabatic integrators, which can take time steps that are substantially larger than the almost-periods of the oscillations.

XIV.1.1 Adiabatic Transformation and Adiabatic Invariants

It comes from the greek $\alpha\delta\iota\alpha\beta\alpha\tau\iota\chi\omicron\varsigma$, “which cannot be crossed”.

... we arrive by analogy to the “adiabatic principle” used in Quantum and then Classical Mechanics. It is based upon the fact that the harmonic oscillator (and other simple dynamical systems as it was found later) submitted to slow variations of its parameters modifies its energy but keeps its action (energy divided by frequency) constant.

As we can see, the path from the word “adiabatic” used in thermodynamics to the above “adiabatic principle” is tortuous and our greek colleagues are certainly puzzled by sentences such as “the changes in the adiabatic invariant due to [...] crossing” which we shall use later.

(J. Henrard 1993)

We consider the linear differential equation

$$\dot{y}(t) = \frac{1}{\varepsilon} Z(t) y(t), \quad (1.1)$$

where $Z(t)$ is a real skew-symmetric (or complex skew-hermitian) matrix-valued function with time derivatives bounded independently of the small parameter ε . In quantum dynamics such equations arise with $Z(t) = -iH(t)$, where the real symmetric (or hermitian) matrix $H(t)$ represents the quantum Hamiltonian operator in a discrete-level Schrödinger equation. We will also encounter real equations of this type in the treatment of oscillatory classical mechanical systems with time-dependent frequencies. Solutions oscillate with almost-periods $\sim \varepsilon$, while the system matrix changes on a slower time scale ~ 1 .

Transforming the Problem. We begin by looking for a time-dependent linear transformation

$$\eta(t) = T_\varepsilon(t)y(t), \quad (1.2)$$

taking the system to the form

$$\dot{\eta}(t) = S_\varepsilon(t) \eta(t) \quad \text{with} \quad S_\varepsilon = \dot{T}_\varepsilon T_\varepsilon^{-1} + \frac{1}{\varepsilon} T_\varepsilon Z T_\varepsilon^{-1}, \quad (1.3)$$

which is chosen such that $S_\varepsilon(t)$ is of smaller norm than the matrix $\frac{1}{\varepsilon} Z(t)$ of (1.1).

Remark 1.1. A first idea is to freeze $Z(t) \approx Z_*$ over a time step and to choose the transformation

$$T_\varepsilon(t) = \exp\left(-\frac{t}{\varepsilon} Z_*\right) \quad \text{yielding} \quad S_\varepsilon(t) = \frac{1}{\varepsilon} \exp\left(-\frac{t}{\varepsilon} Z_*\right) (Z(t) - Z_*) \exp\left(\frac{t}{\varepsilon} Z_*\right).$$

This matrix function $S_\varepsilon(t)$ is highly oscillatory and bounded in norm by $\mathcal{O}(h/\varepsilon)$ for $|t - t_0| \leq h$, if $Z_* = Z(t_0 + h/2)$. Numerical integrators based on this transformation are given by Lawson (1967) and more recently by Hochbruck & Lubich (1999b), Iserles (2002, 2004), and Degani & Schiff (2003). Reasonable accuracy still requires step sizes $h = \mathcal{O}(\varepsilon)$ in general; see also Exercise 3. In the above papers this transformation has, however, been put to good use in situations where the time derivatives of the matrix in the differential equation have much smaller norm than the matrix itself.

Adiabatic Transformation. In order to obtain a differential equation (1.3) with a uniformly bounded matrix $S_\varepsilon(t)$ we diagonalize

$$Z(t) = U(t) i\Lambda(t) U(t)^*$$

with a real diagonal matrix $\Lambda(t) = \text{diag}(\lambda_j(t))$ and a unitary matrix $U(t) = (u_1(t), \dots, u_n(t))$ of eigenvectors depending smoothly on t (possibly except where eigenvalues cross). We define $\eta(t)$ by the unitary *adiabatic transformation*

$$\eta(t) = \exp\left(-\frac{i}{\varepsilon} \Phi(t)\right) U(t)^* y(t) \quad \text{with} \quad \Phi(t) = \text{diag}(\phi_j(t)) = \int_0^t \Lambda(s) ds, \quad (1.4)$$

which represents the solution in a rotating frame of eigenvectors. Each component of $\eta(t)$ is a coefficient in the eigenbasis representation of $y(t)$ rotated in the complex plane by the negative phase. Such transformations have been in use in quantum mechanics since the work of Born & Fock (1928) on adiabatic invariants in Schrödinger equations, as discussed in the next paragraph. The transformation (1.4) yields a differential equation where the ε -independent skew-hermitian matrix

$$W(t) = \dot{U}(t)^* U(t)$$

is framed by oscillatory diagonal matrices:

$$\dot{\eta}(t) = \exp\left(-\frac{i}{\varepsilon}\Phi(t)\right) W(t) \exp\left(\frac{i}{\varepsilon}\Phi(t)\right) \eta(t). \quad (1.5)$$

Numerical integrators for (1.1) based on the transformation to the differential equation (1.5) with bounded, though highly oscillatory right-hand side, are given by Jahnke & Lubich (2003) and Jahnke (2004a); see Sect. XIV.1.2.

Adiabatic Invariants. Possibly after a time-dependent rephasing of the eigenvectors, $u_k(t) \rightarrow e^{i\alpha_k(t)} u_k(t)$, we can assume that $\dot{u}_k(t)$ is orthogonal to $u_k(t)$ for all t . (This is automatically satisfied if $U(t)$ is a real orthogonal matrix, as is the case for $Z(t) = -iH(t)$ with a real symmetric matrix $H(t)$.) We then have the matrix $W = (w_{jk}) = (\dot{u}_j^* u_k)$ with zero diagonal.

After integration of both sides of the differential equation (1.5) from 0 to t , partial integration of the terms on the right-hand side yields for $j \neq k$ (terms for $j = k$ do not appear since $w_{jj} = 0$)

$$\begin{aligned} & \int_0^t \exp\left(-\frac{i}{\varepsilon}(\phi_j(s) - \phi_k(s))\right) w_{jk}(s) \eta_k(s) ds \\ &= i\varepsilon \exp\left(-\frac{i}{\varepsilon}(\phi_j(s) - \phi_k(s))\right) \frac{w_{jk}(s) \eta_k(s)}{\lambda_j(s) - \lambda_k(s)} \Big|_0^t \\ & \quad - i\varepsilon \int_0^t \exp\left(-\frac{i}{\varepsilon}(\phi_j(s) - \phi_k(s))\right) \frac{d}{ds} \frac{w_{jk}(s) \eta_k(s)}{\lambda_j(s) - \lambda_k(s)} ds. \end{aligned} \quad (1.6)$$

At this point, suppose that the eigenvalues $\lambda_j(t)$ are, for all t , separated from each other by a positive distance δ independent of ε :

$$|\lambda_j(t) - \lambda_k(t)| \geq \delta \quad \text{for all } j \neq k. \quad (1.7)$$

Then the reciprocals of their differences and the coupling matrix $W(t)$ are bounded independently of ε , as are their derivatives. Together with the boundedness of $\dot{\eta}$ as implied by (1.5), this shows

$$\eta(t) = \eta(0) + \mathcal{O}(\varepsilon) \quad \text{for } t \leq \text{Const.} \quad (1.8)$$

This result is a version of the quantum-adiabatic theorem of Born & Fock (1928) which states that the actions $|\eta_j|^2$ (the energy in the j th state, $\langle \eta_j u_j, H \eta_j u_j \rangle =$

$\lambda_j |\eta_j|^2$, divided by the frequency λ_j) remain approximately constant for times $t = \mathcal{O}(1)$. Such functions $I(y, t)$ that satisfy $I(y(t), t) = I(y(0), 0) + \mathcal{O}(\varepsilon)$ for $t = \mathcal{O}(1)$ along every $\mathcal{O}(1)$ -bounded solution $y(t)$ of the differential equation, are called *adiabatic invariants*.

Super-Adiabatic Transformations. Adiabatic invariants are obtained over longer time scales by refining the transformation; see Lenard (1959) and Garrido (1964). Here we show that the transformation matrix T_ε of (1.2) can be constructed such that the matrix S_ε in the transformed differential equation (1.3) is of size $\mathcal{O}(\varepsilon^N)$. Let us make the ansatz of a unitary transformation matrix

$$T_\varepsilon^{(N)} = \exp\left(-\frac{i}{\varepsilon}\Phi\right) \exp(-i\Phi_1) \exp(\varepsilon X_1) \dots \exp(-i\varepsilon^{N-1}\Phi_N) \exp(\varepsilon^N X_N) U^*$$

with real diagonal matrices $\Phi_n(t)$ and complex skew-hermitian matrices $X_n(t)$. We find that $S_\varepsilon = \frac{1}{\varepsilon} T_\varepsilon Z T_\varepsilon^* + \dot{T}_\varepsilon T_\varepsilon^*$ is $\mathcal{O}(\varepsilon)$ if and only if X_1 and $A_1 := \dot{\Phi}_1$ satisfy

$$\frac{1}{\varepsilon} \left(\exp(\varepsilon X_1) iA \exp(-\varepsilon X_1) - iA \right) - iA_1 + W = \mathcal{O}(\varepsilon),$$

or equivalently, if X_1 and A_1 solve the commutator equation

$$[iA, X_1] + iA_1 = W.$$

This is solved by setting iA_1 equal to the diagonal of W and determining the off-diagonal entries $x_{jk}^{(1)}$ of X_1 from the scalar equations

$$i(\lambda_j - \lambda_k) x_{jk}^{(1)} = w_{jk}, \quad j \neq k,$$

which can be done as long as the eigenvalues are separated. The diagonal of X_1 is set to zero. Since W is skew-hermitian, so is X_1 . Similarly we obtain for higher powers of ε the equations

$$[iA, X_n] + iA_n = W_{n-1},$$

where the matrix W_{n-1} contains only previously constructed terms up to index $n-1$ and derivatives up to order n and is skew-hermitian because S_ε is skew-hermitian. In this way we obtain a unitary transformation such that

$$\eta^{(N)}(t) = T_\varepsilon^{(N)}(t) y(t) \quad \text{satisfies} \quad \dot{\eta}^{(N)} = \mathcal{O}(\varepsilon^N).$$

We remark that the above construction of $T_\varepsilon^{(N)}$ is analogous to transformations in Hamiltonian perturbation theory; cf. Sect. X.2.

The differential equation (1.1) thus has adiabatic invariants over times $\mathcal{O}(\varepsilon^{-N})$ for arbitrary $N \geq 1$, and in fact even over exponentially long time intervals $t = \mathcal{O}(e^{c/\varepsilon})$ if the functions have a bounded analytic extension to a complex strip, as is shown by Joye & Pfister (1993) and Nenciu (1993). The leading term in the exponentially small deviation of $|\eta_j^{(N)}(t)|^2$ in the optimally truncated super-adiabatic basis has been rigorously made explicit by Betz & Teufel (2005a, 2005b), proving a conjecture by Berry (1990).

Avoided Crossing of Eigenvalues and Non-Adiabatic Transitions. To illustrate the effects of a violation of the separation condition (1.7), we consider the generic two-dimensional example studied by Zener (1932), with the matrix

$$Z(t) = -i \begin{pmatrix} t & \delta \\ \delta & -t \end{pmatrix}, \quad (1.9)$$

which has the eigenvalues $\pm i\sqrt{t^2 + \delta^2}$. The minimal distance of the eigenvalues is 2δ at $t = 0$. For $\delta = \mathcal{O}(\sqrt{\varepsilon})$ the adiabatic invariance (1.8) is no longer valid, and η can undergo $\mathcal{O}(1)$ changes in an $\mathcal{O}(\delta)$ neighbourhood of $t = 0$: a *non-adiabatic transition* in physical terminology. The changes in the adiabatic invariant due to the avoided crossing of eigenvalues are illustrated in Fig. 1.1 and can be explained as follows.

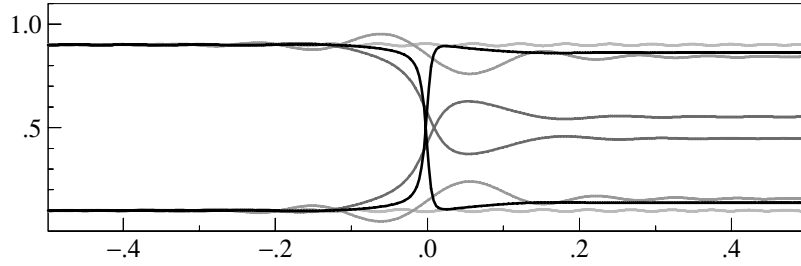


Fig. 1.1. Non-adiabatic transition: $|\eta_1(t)|^2$ and $|\eta_2(t)|^2$ as function of t for $\varepsilon = 0.01$ and $\delta = 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}$ (increasing darkness)

Near the avoided crossing, a new time scale $\tau = t/\delta$ is appropriate. The decomposition $Z(t) = U(t)i\Lambda(t)U(t)^T$ of the matrix yields

$$\begin{aligned} U(t) &= \tilde{U}(\tau) = \begin{pmatrix} \cos \alpha(\tau) & -\sin \alpha(\tau) \\ \sin \alpha(\tau) & \cos \alpha(\tau) \end{pmatrix}, \\ \Lambda(t)/\delta &= \tilde{\Lambda}(\tau) = \begin{pmatrix} -\sqrt{\tau^2 + 1} & 0 \\ 0 & \sqrt{\tau^2 + 1} \end{pmatrix}, \end{aligned}$$

with $\alpha(\tau) = \frac{\pi}{4} - \frac{1}{2} \arctan(\tau)$. We introduce the rescaled matrices

$$\begin{aligned} \tilde{\Phi}(\tau) &= \int_0^\tau \tilde{\Lambda}(\sigma) d\sigma = \Phi(t)/\delta^2, \\ \tilde{W}(\tau) &= \left(\frac{d}{d\tau} \tilde{U}(\tau)^T \right) \tilde{U}(\tau) = \frac{1}{2(\tau^2 + 1)} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \delta \cdot W(t). \end{aligned}$$

Note that the entries of $W(t)$ have a sharp peak of height $(2\delta)^{-1}$ at $t = 0$. The rescaled function $\tilde{\eta}(\tau) = \eta(t)$ is a solution of the differential equation

$$\frac{d}{d\tau} \tilde{\eta}(\tau) = \exp \left(-\frac{i\delta^2}{\varepsilon} \tilde{\Phi}(\tau) \right) \tilde{W}(\tau) \exp \left(\frac{i\delta^2}{\varepsilon} \tilde{\Phi}(\tau) \right) \tilde{\eta}(\tau).$$

For $\delta^2 \leq \varepsilon$ and $|\tau| = |t/\delta| \leq 1$, the matrix on the right-hand side is bounded of norm ~ 1 and has bounded derivatives with respect to τ . The function $\tilde{\eta}(\tau)$ therefore changes its value by an amount of size $\mathcal{O}(1)$ in the interval $|\tau| \leq 1$. We also note that any numerical integrator using piecewise polynomial approximations of $W(t)$ and hence of $\tilde{W}(\tau)$ must take step sizes $\Delta\tau = h/\delta \ll 1$, i.e., $h \ll \delta$. On the other hand, the rescaling shows that the number of time steps needed to resolve the non-adiabatic transition up to a specified accuracy is independent of δ .

XIV.1.2 Adiabatic Integrators

We discuss symmetric long-time-step integrators for the rotating-frame differential equation (1.5) that describes skew-hermitian systems in adiabatic variables. The construction follows Jahnke & Lubich (2003) and Jahnke (2004a); see also Lorenz, Jahnke & Lubich (2005).

First-Order Integrators. We consider the differential equation (1.5) and integrate both sides from t_n to $t_{n+1} = t_n + h$:

$$\eta(t_{n+1}) = \eta(t_n) + \int_{t_n}^{t_{n+1}} \exp\left(-\frac{i}{\varepsilon}\Phi(s)\right) W(s) \exp\left(\frac{i}{\varepsilon}\Phi(s)\right) \eta(s) ds, \quad (1.10)$$

where $W(t)$ is an ε -independent matrix, continuously differentiable in t , and the real diagonal matrix of phases $\Phi(t)$ is given as the integral of $\Lambda(t) = \text{diag}(\lambda_j(t))$. In the applications, $W(t)$ and $\Phi(t)$ are not given explicitly, but need to be computed using numerical differentiation and integration, respectively. For simplicity, we here ignore this approximation and consider W , Φ , Λ as given time-dependent functions.

Since η and W have bounded derivatives, the following averaged version of the implicit midpoint rule has a local error of $\mathcal{O}(h^2)$ uniformly in ε :¹

$$\eta_{n+1} = \eta_n + \int_{t_n}^{t_{n+1}} \exp\left(-\frac{i}{\varepsilon}\Phi(s)\right) W(t_{n+1/2}) \exp\left(\frac{i}{\varepsilon}\Phi(s)\right) ds \frac{1}{2}(\eta_{n+1} + \eta_n). \quad (1.11)$$

The problem then remains to compute the oscillatory integral. The integrand can be rewritten as

$$E(\Phi(s)) \bullet W(t_{n+1/2}),$$

where \bullet denotes the entrywise product of matrices and

$$E(\Phi) = (e_{jk}) \quad \text{with} \quad e_{jk} = \exp\left(-\frac{i}{\varepsilon}(\phi_j - \phi_k)\right).$$

With a linear phase approximation (of an error $\mathcal{O}(h^2)$)

$$\Phi(t_{n+1/2} + \theta h) \approx \Phi(t_{n+1/2}) + \theta h \Lambda(t_{n+1/2}),$$

¹ Because of the oscillatory integrals, the local error is not $\mathcal{O}(h^3)$ as might at first glance be expected for a symmetric method.

the integral is approximated by

$$h E(\Phi(t_{n+1/2})) \bullet \mathcal{I}(t_{n+1/2}) \bullet W(t_{n+1/2})$$

where $\mathcal{I}(t)$ is the matrix of integrated exponentials with entries (we omit the argument t)

$$\mathcal{I}_{jk} = \int_{-1/2}^{1/2} \exp\left(-\frac{i\theta h}{\varepsilon}(\lambda_j - \lambda_k)\right) d\theta = \text{sinc}\left(\frac{h}{2\varepsilon}(\lambda_j - \lambda_k)\right).$$

The error in the integral approximation comes solely from the linear phase approximation and is bounded by $\mathcal{O}(h \cdot \frac{h^2}{\varepsilon} \cdot \frac{\varepsilon}{h}) = \mathcal{O}(h^2)$ if the λ_j are separated, because then the integral \mathcal{I}_{jk} is of size $\mathcal{O}(\frac{\varepsilon}{h})$. We thus obtain the following *averaged implicit midpoint rule* with a local error of $\mathcal{O}(h^2)$ uniformly in ε :

$$\eta_{n+1} = \eta_n + h \left(E(\Phi(t_{n+1/2})) \bullet \mathcal{I}(t_{n+1/2}) \bullet W(t_{n+1/2}) \right) \frac{1}{2}(\eta_{n+1} + \eta_n). \quad (1.12)$$

An analogue of the explicit midpoint rule is similarly constructed, and from the Magnus series (IV.7.5) of the solution we obtain the following *averaged exponential midpoint rule*, again with an $\mathcal{O}(h^2)$ local error uniformly in ε :

$$\eta_{n+1} = \exp\left(h E(\Phi(t_{n+1/2})) \bullet \mathcal{I}(t_{n+1/2}) \bullet W(t_{n+1/2})\right) \eta_n. \quad (1.13)$$

For skew-hermitian $W(t)$, also the matrix in (1.12) and (1.13) is skew-hermitian, and hence both of the above integrators preserve the Euclidean norm of η exactly. We summarize the local error bounds for these methods under conditions that include the case of an avoided crossing of eigenvalues.

Theorem 1.2 (Local Error). *Suppose that for $t_0 \leq t \leq t_0 + h$ and all j, k ,*

$$|\lambda_j(t) - \lambda_k(t)| \geq \delta, \quad |\dot{\lambda}_j(t)| \leq C_0, \quad \|W(t)\| \leq \frac{C_1}{\delta}, \quad \|\dot{W}(t)\| \leq \frac{C_2}{\delta^2}$$

with $\delta > 0$. Then, the local error of methods (1.12) and (1.13) is bounded by

$$\|\eta_1 - \eta(t_0 + h)\| \leq C \frac{h^2}{\delta^2} \|\eta_0\|.$$

The constant C is independent of h, ε, δ .

Proof. The result is obtained with the arguments and approximation estimates given above, taking in addition account of the dependence on δ . \square

The local error contains smooth, non-oscillatory components which accumulate to a global error $\eta_n - \eta(t_n) = \mathcal{O}(h)$ on bounded intervals if the eigenvalues remain well separated. Using that in this case η is constant up to $\mathcal{O}(\varepsilon)$, this error bound can be improved to $\mathcal{O}(\min\{\varepsilon, h\})$. The integrators thus do not resolve the $\mathcal{O}(\varepsilon)$ oscillations in η for large step sizes $h \geq \varepsilon$, but like in Jahnke & Lubich (2003)

they can be combined with a (symmetric and scaling-invariant) adaptive step size strategy such that the methods follow the non-adiabatic transitions through avoided crossings of eigenvalues with small steps and take large steps elsewhere.

We here consider applying an *integrating reversible step size controller* as in Sect. VIII.3.2 with the step size density function

$$\sigma(t) = (\|W(t)\|^2 + \alpha^2)^{-1/2}$$

for a parameter α that can be interpreted as the ratio of the accuracy parameter and the maximum admissible step size. Choosing the Frobenius norm $\|W\| = (\text{trace } W^T W)^{1/2}$, we then obtain the following version of Algorithm VIII.3.4, where μ is the accuracy parameter and

$$G(t) = -\frac{\dot{\sigma}(t)}{\sigma(t)} = (\|W(t)\|^2 + \alpha^2)^{-1} \text{trace } (\dot{W}(t)^T W(t)).$$

Set $z_0 = 1/\sigma(t_0)$ and, for $n \geq 0$,

$$\begin{aligned} z_{n+1/2} &= z_n + \frac{\mu}{2} G(t_n) \\ h_{n+1/2} &= \mu / z_{n+1/2} \\ t_{n+1} &= t_n + h_{n+1/2} \\ \eta_n &\mapsto \eta_{n+1} \quad \text{by (1.12) or (1.13) with step size } h_{n+1/2} \\ z_{n+1} &= z_{n+1/2} + \frac{\mu}{2} G(t_{n+1}). \end{aligned} \tag{1.14}$$

We remark that the schemes (1.12) and (1.13) can be modified such that they use evaluations at t_n and t_{n+1} instead of $t_{n+1/2}$ (Exercise 6).

Applying the above algorithm with accuracy parameter $\mu = 0.01$ and $\alpha = 0.1$ to the problem of Fig. 1.1 with $\varepsilon = 0.01$ and $\delta = 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}$ yields the step size sequences shown in Fig. 1.2. In each case the error at the end-point $t = 1$ was between $0.5 \cdot 10^{-3}$ and $2 \cdot 10^{-3}$.

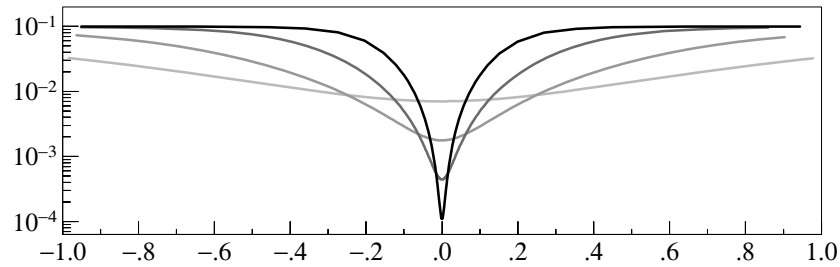


Fig. 1.2. Non-adiabatic transition: step sizes as function of t for $\varepsilon = 0.01$ and $\delta = 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}$ (increasing darkness)

Second-Order Integrators. The $\mathcal{O}(\varepsilon)$ oscillations in η are resolved with step sizes up to $h = \mathcal{O}(\sqrt{\varepsilon})$ for methods that give $\mathcal{O}(h^2)$ accuracy uniformly in ε . Such

methods require a quadratic phase approximation, and one needs further terms obtained from reinserting $\eta(s)$ under the integral in (1.10) once again by the same formula, thus yielding terms with iterated integrals (this procedure is known as the *Neumann* or *Peano* or *Dyson* expansion in different communities, cf. Iserles 2004), or by including the first commutator in the *Magnus* expansion (IV.7.5). Symmetric second-order methods of both types are constructed by Jahnke (2004a).

Care must be taken in computing the arising oscillatory integrals. Iserles (2004) proposes and analyses Filon quadrature (after Filon, 1928), which is applicable when the moments, i.e., the integrals over products of oscillatory exponentials and polynomials, are known analytically. This is not the case, however, for all of the integrals appearing in the second-order methods. The alternative chosen by Jahnke (2004a) is to use an expansion technique based on partial integration. The idea can be illustrated on an integral such as

$$\int_0^1 \exp\left(\frac{i\alpha\theta h}{\varepsilon}\right) \cdot \exp\left(\frac{i\beta\theta^2 h^2}{\varepsilon}\right) d\theta$$

with $\alpha \neq 0$. Partial integration that integrates the first factor and differentiates the second factor yields a boundary term and again an integral of the same type, but now with an additional factor $\mathcal{O}\left(\frac{\varepsilon}{h} \cdot \frac{h^2}{\varepsilon}\right) = \mathcal{O}(h)$. Using this technique repeatedly in the oscillatory integrals appearing in the second-order methods permits to approximate all of them up to $\mathcal{O}(h^3)$ as needed. We refer to Jahnke (2004a) for the precise formulation and error analysis of these second-order methods, which are complicated to formulate, but do not require substantially more computational work than the first-order methods described above, and just the same number of matrix evaluations.

Higher-Order Integrators. Integrators of general order $p \geq 1$ are obtained with a phase approximation by polynomials of degree p and by including all terms of the Neumann or Magnus expansion for (1.5) with up to p -fold integrals.

XIV.2 Mechanical Systems with Time-Dependent Frequencies

We study oscillatory mechanical systems with explicitly time-dependent frequencies, where the time-dependent Hamiltonian is

$$H(p, q, t) = \frac{1}{2} p^T M(t)^{-1} p + \frac{1}{2\varepsilon^2} q^T A(t) q + U(q, t) \quad (2.1)$$

with a positive definite mass matrix $M(t)$ and a positive semi-definite stiffness matrix $A(t)$ of constant rank whose derivatives are bounded independently of ε . Such a Hamiltonian describes oscillations in a mechanical system that at the same time exerts a driven motion on a slower time scale. We consider motions of bounded energy:

$$H(p(t), q(t), t) \leq \text{Const.} \quad (2.2)$$

We transform (2.1) to a more amenable form by a series of linear time-dependent canonical coordinate transforms. The transformations turn the equations of motion into a form that approximately separates the time scales. This makes the problem more accessible to numerical discretization with large time steps and to the error analysis of multiple time-stepping methods applied directly to (2.1) in the originally given coordinates.

XIV.2.1 Canonical Transformation to Adiabatic Variables

By a series of canonical time-dependent linear transformations, which can all be done numerically with standard linear algebra routines, we now take the Hamiltonian system (2.1) to a form from which adiabatic invariants can be read off and which will serve as the base camp for both the construction and error analysis of numerical methods.

We introduce the energy E as the conjugate variable to time t and extend the Hamiltonian to

$$\hat{H}(p, E, q, t) = H(p, q, t) + E. \quad (2.3)$$

The canonical equations of motion are then (the gradient ∇ refers only to q)

$$\begin{aligned} \dot{p} &= -\frac{1}{\varepsilon^2} A(t)q - \nabla U(q, t) \\ \dot{q} &= M(t)^{-1}p \end{aligned}$$

along with $\dot{E} = -\partial H / \partial t$ and $\dot{t} = 1$.

Transforming the Mass Matrix into the Identity Matrix. We change variables such that the mass matrix $M(t)$ in the kinetic energy part is replaced by the identity. With a smooth factorization

$$M(t)^{-1} = C(t)C(t)^T, \quad (2.4)$$

e.g., from a Cholesky decomposition of $M(t)$, we transform to variables (\tilde{q}, \tilde{t}) by

$$q = C(t)\tilde{q}, \quad t = \tilde{t}.$$

Then, the conjugate momenta are given by (see Example VI.5.2)

$$\begin{pmatrix} \tilde{p} \\ \tilde{E} \end{pmatrix} = \begin{pmatrix} C & \dot{C}\tilde{q} \\ 0 & 1 \end{pmatrix}^T \begin{pmatrix} p \\ E \end{pmatrix} = \begin{pmatrix} C^T p \\ \tilde{q}^T \dot{C}^T p + E \end{pmatrix}.$$

With the transformed matrix $\tilde{A} = C^T A C$, the Hamiltonian $\tilde{H}(\tilde{p}, \tilde{E}, \tilde{q}, \tilde{t}) = \hat{H}(p, E, q, t)$ in the new variables then takes the form (we omit all tildes)

$$H(p, E, q, t) = \frac{1}{2} p^T p + \frac{1}{2\varepsilon^2} q^T A(t)q - q^T \dot{C}(t)^T C(t)^{-T} p + U(C(t)q, t) + E. \quad (2.5)$$

Diagonalizing the Stiffness Matrix. We diagonalize the matrix $A(t)$ in (2.5),

$$A(t) = Q(t) \begin{pmatrix} 0 & 0 \\ 0 & \Omega(t)^2 \end{pmatrix} Q(t)^T \quad (2.6)$$

with the diagonal matrix $\Omega(t) = \text{diag}(\omega_j(t))$ of frequencies and an orthogonal matrix $Q(t)$, which depends smoothly on t if the frequencies remain separated. The matrix $Q(t)$ can be obtained as the product

$$Q(t) = Q_0(t) \begin{pmatrix} I & 0 \\ 0 & Q_*(t) \end{pmatrix}, \quad (2.7)$$

where the transformation with $Q_0(t)$ takes $A(t)$ to the block-diagonal form

$$A(t) = Q_0(t) \begin{pmatrix} 0 & 0 \\ 0 & A_*(t) \end{pmatrix} Q_0(t)^T$$

and $Q_*(t)$ diagonalizes $A_*(t)$. The effect of an avoided crossing of frequencies is localized to $Q_*(t)$, which then can have large derivatives, whereas those of $Q_0(t)$ remain moderately bounded. The transformation

$$q = Q(t)\hat{q}, \quad t = \hat{t}$$

with the conjugate momenta

$$\hat{p} = Q(t)^T p, \quad \hat{E} = \hat{q}^T \dot{Q}(t)^T p + E$$

yields the Hamiltonian in the new variables $(\hat{p}, \hat{E}, \hat{q}, \hat{t})$ as (we omit all hats)

$$H = \frac{1}{2} p^T p + \frac{1}{2\varepsilon^2} q^T \begin{pmatrix} 0 & 0 \\ 0 & \Omega(t)^2 \end{pmatrix} q + q^T K(t) p + U(C(t)Q(t)q, t) + E \quad (2.8)$$

with

$$K = \begin{pmatrix} K_{00} & K_{01} \\ K_{10} & K_{11} \end{pmatrix} = Q^T \dot{Q} - Q^T \dot{C}^T C^{-T} Q.$$

We decompose also

$$p = \begin{pmatrix} p_0 \\ p_1 \end{pmatrix}, \quad q = \begin{pmatrix} q_0 \\ q_1 \end{pmatrix}$$

according to the blocks in (2.6) and refer to q_0 and q_1 (p_0 and p_1) as the *slow* and *fast* positions (slow and fast momenta), respectively. With the energy bound (2.2) we have

$$p_1 = \mathcal{O}(1), \quad q_1 = \mathcal{O}(\varepsilon). \quad (2.9)$$

Rescaling Positions and Momenta. We transform

$$q_0 = \check{q}_0, \quad q_1 = \varepsilon^{1/2} \Omega^{-1/2} \check{q}_1, \quad t = \check{t}$$

with the conjugate momenta

$$\check{p}_0 = p_0, \quad \check{p}_1 = \varepsilon^{1/2} \Omega^{-1/2} p_1, \quad \check{E} = -\frac{1}{2} \check{q}_1^T \varepsilon^{1/2} \Omega^{-3/2} \dot{\check{q}}_1 + E.$$

In the new variables, the Hamiltonian becomes (we omit the hačeks on all variables)

$$\begin{aligned} H &= \frac{1}{2} p_0^T p_0 + \frac{1}{2\varepsilon} p_1^T \Omega(t) p_1 + \frac{1}{2\varepsilon} q_1^T \Omega(t) q_1 \\ &\quad + q^T \check{K}(t) p + U(T(t)q, t) + E \end{aligned} \quad (2.10)$$

with

$$\begin{aligned} \check{K} &= \begin{pmatrix} K_{00} & \varepsilon^{-1/2} K_{01} \Omega^{1/2} \\ \varepsilon^{1/2} \Omega^{-1/2} K_{10} & \Omega^{-1/2} K_{11} \Omega^{1/2} + \frac{1}{2} \Omega^{-1} \dot{\Omega} \end{pmatrix} \\ T &= \left(T_0 \mid \varepsilon^{1/2} T_1 \right) = \begin{pmatrix} T_{00} & \varepsilon^{1/2} T_{01} \\ T_{10} & \varepsilon^{1/2} T_{11} \end{pmatrix} = CQ \begin{pmatrix} I & 0 \\ 0 & \varepsilon^{1/2} \Omega^{-1/2} \end{pmatrix}. \end{aligned}$$

Eliminating the Singular Block. We next remove the $\mathcal{O}(\varepsilon^{-1/2})$ off-diagonal block in \check{K} by the canonical transformation

$$-p_1 = -\bar{p}_1 + \varepsilon^{1/2} \Omega^{-1/2} K_{01}^T q_0, \quad q_0 = \bar{q}_0, \quad t = \bar{t}$$

with the conjugate variables

$$\bar{q}_1 = q_1, \quad \bar{p}_0 = p_0 + \varepsilon^{1/2} K_{01} \Omega^{-1/2} q_1, \quad \bar{E} = E + \varepsilon^{1/2} q_0^T \frac{d}{dt} (K_{01} \Omega^{-1/2}) q_1.$$

In these coordinates, the Hamiltonian takes the form (we omit all bars)

$$\begin{aligned} H &= \frac{1}{2} p_0^T p_0 + \frac{1}{2\varepsilon} p_1^T \Omega(t) p_1 + \frac{1}{2\varepsilon} q_1^T \Omega(t) q_1 \\ &\quad + q^T L(t) p + \frac{1}{2} q^T S(t) q + U(T(t)q, t) + E \end{aligned} \quad (2.11)$$

with the lower block-triangular matrix

$$\begin{aligned} L &= \begin{pmatrix} L_{00} & 0 \\ \varepsilon^{1/2} L_{10} & L_{11} \end{pmatrix} \\ &= \begin{pmatrix} K_{00} & 0 \\ \varepsilon^{1/2} \Omega^{-1/2} (K_{10} + K_{01}^T) & \Omega^{-1/2} K_{11} \Omega^{1/2} + \frac{1}{2} \Omega^{-1} \dot{\Omega} \end{pmatrix} \end{aligned}$$

and the symmetric matrix

$$S = \begin{pmatrix} S_{00} & \varepsilon^{1/2} S_{01} \\ \varepsilon^{1/2} S_{10} & \varepsilon S_{11} \end{pmatrix},$$

where

$$\begin{aligned} S_{00} &= -K_{01} K_{01}^T, \\ S_{01} &= S_{10}^T = -K_{00} K_{01} \Omega^{-1/2} \\ &\quad - K_{01} \Omega^{-1/2} (\Omega^{1/2} K_{11}^T \Omega^{-1/2} + \frac{1}{2} \Omega^{-1} \dot{\Omega}) - \frac{d}{dt} (K_{01} \Omega^{-1/2}), \\ S_{11} &= \Omega^{-1/2} (-K_{10} K_{01} - K_{01}^T K_{10}^T + K_{01}^T K_{01}) \Omega^{-1/2}. \end{aligned}$$

We note that with the energy bound (2.2) we now have

$$p_1 = \mathcal{O}(\varepsilon^{1/2}), \quad q_1 = \mathcal{O}(\varepsilon^{1/2}). \quad (2.12)$$

Equations of Motion. The differential equations now take the form

$$\begin{aligned} \dot{p}_0 &= f_0(p, q, t) \\ \dot{q}_0 &= p_0 + g_0(q, t) \\ \begin{pmatrix} \dot{p}_1 \\ \dot{q}_1 \end{pmatrix} &= \frac{1}{\varepsilon} \begin{pmatrix} 0 & -\Omega(t) \\ \Omega(t) & 0 \end{pmatrix} \begin{pmatrix} p_1 \\ q_1 \end{pmatrix} + \begin{pmatrix} f_1(p, q, t) \\ g_1(q, t) \end{pmatrix} \end{aligned} \quad (2.13)$$

with the functions bounded uniformly in ε ,

$$\begin{pmatrix} f_0 \\ f_1 \end{pmatrix} = -L(t)p - S(t)q - T(t)^T \nabla U(T(t)q, t), \quad \begin{pmatrix} g_0 \\ g_1 \end{pmatrix} = L(t)^T q.$$

The matrix in the system is diagonalized by a constant unitary matrix: with

$$\Gamma = \frac{1}{\sqrt{2}} \begin{pmatrix} I & I \\ -iI & iI \end{pmatrix} \quad (2.14)$$

we have

$$\begin{pmatrix} 0 & -\Omega(t) \\ \Omega(t) & 0 \end{pmatrix} = \Gamma \begin{pmatrix} i\Omega(t) & 0 \\ 0 & -i\Omega(t) \end{pmatrix} \Gamma^*. \quad (2.15)$$

Remark. Action-angle variables $p_{1,j} = \sqrt{a_j} \cos \theta_j$, $q_{1,j} = \sqrt{a_j} \sin \theta_j$ for the harmonic oscillators would now put the Hamiltonian into the form $H = \frac{1}{\varepsilon} \omega(t) \cdot a + G(a, \theta, p_0, q_0, t)$, which could be studied further using averaging techniques, that is, using coordinate transforms that reduce the dependence on the angles in the Hamiltonian; see Neishtadt (1984) for averaging out up to an exponentially small remainder in the case of a single high frequency. The first-order averaging transform might be done numerically (cf. the formulas in Sect. XII.2), but the higher-order transforms involve increasingly higher derivatives of the functions involved and therefore become impractical from the numerical viewpoint. For systems with several frequencies the averaging transforms require multi-dimensional integrals which are

expensive to compute. For our numerical purposes we therefore continue differently, adapting the adiabatic transformation of Sect. XIV.1.1.

The System in Adiabatic Variables. Let the diagonal phase matrix be given as

$$\Phi(t) = \int_{t_0}^t \Lambda(s) ds \quad \text{with} \quad \Lambda(t) = \begin{pmatrix} \Omega(t) & 0 \\ 0 & -\Omega(t) \end{pmatrix}.$$

Our final transformation follows (1.4) and sets

$$\eta = \varepsilon^{-1/2} \exp\left(-\frac{i}{\varepsilon} \Phi(t)\right) \Gamma^* \begin{pmatrix} p_1 \\ q_1 \end{pmatrix}. \quad (2.16)$$

The factor $\varepsilon^{-1/2}$ is chosen for convenience so that (2.12) implies

$$\eta = \mathcal{O}(1). \quad (2.17)$$

We remark that up to now all transformations were invariant under rescaling $\varepsilon \rightarrow \sigma\varepsilon$ and $A(t) \rightarrow \sigma^2 A(t)$, but here we have chosen to give up this invariance in favour of (2.17). Note that η is of the form

$$\eta = \varepsilon^{-1/2} \Gamma^* \begin{pmatrix} \pi \\ \rho \end{pmatrix} = \frac{\varepsilon^{-1/2}}{\sqrt{2}} \begin{pmatrix} \pi + i\rho \\ \pi - i\rho \end{pmatrix} \quad (2.18)$$

with real vectors π, ρ satisfying

$$\pi + i\rho = \exp\left(-\frac{i}{\varepsilon} \int_{t_0}^t \Omega(s) ds\right) (p_1 + iq_1). \quad (2.19)$$

We denote the inverse transform as

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \varepsilon^{1/2} \begin{pmatrix} P_1(t) \\ Q_1(t) \end{pmatrix} \eta \quad \text{with} \quad \begin{pmatrix} P_1(t) \\ Q_1(t) \end{pmatrix} = \Gamma \exp\left(\frac{i}{\varepsilon} \Phi(t)\right). \quad (2.20)$$

Together with $e = E + \frac{1}{2\varepsilon} p_1^T \Omega(t) p_1 + \frac{1}{2\varepsilon} q_1^T \Omega(t) q_1$ and unaltered p_0, q_0, t this yields a canonical transformation $(p_0, \pi, e, q_0, \rho, t) \mapsto (p_0, p_1, E, q_0, q_1, t)$. The Hamiltonian reads in these variables

$$H = \frac{1}{2} p_0^T p_0 + q^T L(t) p + \frac{1}{2} q^T S(t) q + U(T(t) q, t) + e,$$

where on the right-hand side the components p_1, q_1 are expressed in terms of η and π, ρ by (2.20) and (2.18). The equations of motion now become

$$\begin{aligned} \dot{p}_0 &= f_0(p, q, t) \\ \dot{q}_0 &= p_0 + g_0(q, t) \\ \dot{\eta} &= \varepsilon^{-1/2} \exp\left(-\frac{i}{\varepsilon} \Phi(t)\right) \Gamma^* \begin{pmatrix} f_1(p, q, t) \\ g_1(q, t) \end{pmatrix} \end{aligned}$$

with p_1, q_1 expressed in terms of η by (2.20). Written out, the differential equations for p_0, q_0 read

$$\begin{aligned}\dot{p}_0 &= -L_{00}p_0 - S_{00}q_0 - T_0^T \nabla U(T_0 q_0, t) - \varepsilon S_{01}Q_1\eta \\ &\quad - T_0^T \left(\nabla U(T_0 q_0 + \varepsilon T_1 Q_1 \eta, t) - \nabla U(T_0 q_0, t) \right) \\ \dot{q}_0 &= p_0 + L_{00}^T q_0 + \varepsilon L_{10}^T Q_1 \eta.\end{aligned}\quad (2.21)$$

The matrix multiplying η after substituting the expressions f_1 and g_1 in the differential equation for η becomes, apart from the oscillatory exponentials,

$$\begin{aligned}W &= \Gamma^* \begin{pmatrix} -L_{11} & -\varepsilon S_{11} \\ 0 & L_{11}^T \end{pmatrix} \Gamma \\ &= -\frac{1}{2} \begin{pmatrix} L_{11} - L_{11}^T & L_{11} + L_{11}^T \\ L_{11} + L_{11}^T & L_{11} - L_{11}^T \end{pmatrix} - \frac{i\varepsilon}{2} \begin{pmatrix} -S_{11} & S_{11} \\ -S_{11} & S_{11} \end{pmatrix},\end{aligned}\quad (2.22)$$

which has a diagonal of size $\mathcal{O}(\varepsilon)$. The equation for η then reads

$$\begin{aligned}\dot{\eta} &= \exp\left(-\frac{i}{\varepsilon}\Phi(t)\right) W(t) \exp\left(\frac{i}{\varepsilon}\Phi(t)\right) \eta \\ &\quad - P_1^* \left(L_{10}p_0 + S_{10}q_0 + T_1^T \nabla U(T_0 q_0 + \varepsilon T_1 Q_1 \eta, t) \right).\end{aligned}\quad (2.23)$$

The matrix multiplying η is bounded independently of ε , but highly oscillatory. Note that the coordinate transforms leading to (2.21), (2.23) are linear and can be carried out by standard numerical linear algebra routines.

Adiabatic Invariants. We suppose that the eigenfrequencies $\omega_j(t)$ remain separated and bounded away from 0: there are $\delta > 0$ and $c > 0$ such that for any pair $\omega_j(t)$ and $\omega_k(t)$ with $j \neq k$ ($j, k = 1, \dots, m$), the lower bounds

$$|\omega_j(t) - \omega_k(t)| \geq \delta, \quad \omega_j(t) \geq c \quad (2.24)$$

hold for all t under consideration. Under condition (2.24) the right-hand side $r(t)$ in the differential equation for η consists only of oscillatory terms, up to $\mathcal{O}(\varepsilon)$. (No smooth terms larger than $\mathcal{O}(\varepsilon)$ arise because the matrix W has a diagonal of size $\mathcal{O}(\varepsilon)$.) It then follows by partial integration that

$$\int_0^t r(s) ds = \mathcal{O}(\varepsilon) \quad \text{for } t \leq \text{Const.}, \quad (2.25)$$

and as in (1.6) we then obtain

$$\eta(t) = \eta(0) + \mathcal{O}(\varepsilon) \quad \text{for } t \leq \text{Const.} \quad (2.26)$$

The functions defined by

$$I_j = |\eta_j|^2 \quad (j = 1, \dots, m) \quad (2.27)$$

are thus adiabatic invariants:

$$I_j(t) = I_j(0) + \mathcal{O}(\varepsilon) \quad \text{for } t \leq \text{Const.} \quad (2.28)$$

Starting from a Hamiltonian system (2.1), where the mass matrix equals the identity and the stiffness matrix is already diagonal, we find that I_j is the action (energy divided by frequency)

$$I_j(t) = \frac{1}{\omega_j(t)} \left(\frac{1}{2} p_j(t)^2 + \frac{\omega_j(t)^2}{2\varepsilon^2} q_j(t)^2 \right),$$

which for a constant frequency ω_j becomes a constant multiple of the oscillatory energy considered in Sect. XIII.9.

The Slow Limit System. As $\varepsilon \rightarrow 0$, the evolution of the slow variables p_0, q_0 is governed by the equations

$$\begin{aligned} \dot{p}_0 &= -L_{00}(t)p_0 - S_{00}(t)q_0 - T_0(t)^T \nabla U(T_0(t)q_0, t) \\ \dot{q}_0 &= p_0 + L_{00}(t)^T q_0 \end{aligned} \quad (2.29)$$

which is the system with the time-dependent Hamiltonian

$$H_0(p_0, q_0, t) = \frac{1}{2} p_0^T p_0 + q_0^T L_{00}(t) p_0 + \frac{1}{2} q_0^T S_{00}(t) q_0 + U(T_0(t)q_0, t).$$

We conclude this subsection with a simple illustration of the above procedure.

Example 2.1 (Harmonic oscillator with slowly varying frequency). For the scalar second-order differential equation

$$\ddot{q} + \frac{\omega(t)^2}{\varepsilon^2} q = 0,$$

where $\omega(t)$ is bounded away from 0 and has a derivative bounded independently of ε , the above transformations simplify considerably. The Hamiltonian in the original variables is already of the form

$$H = \frac{1}{2} p^2 + \frac{1}{2} \frac{\omega(t)^2}{\varepsilon^2} q^2,$$

and hence the first two transformations are not needed at all, and there are no slow variables p_0, q_0 . The rescaling transformation yields the Hamiltonian (2.10) in the form

$$H = \frac{\omega(t)}{2\varepsilon} \dot{p}^2 + \frac{\omega(t)}{2\varepsilon} \dot{q}^2 + \frac{1}{2} \frac{\dot{\omega}(t)}{\omega(t)} \dot{p} \dot{q}.$$

With the adiabatic transformation (2.19) we thus represent the solution as

$$\sqrt{\frac{\varepsilon}{\omega(t)}} \dot{q}(t) + i \sqrt{\frac{\omega(t)}{\varepsilon}} q(t) = \exp\left(\frac{i}{\varepsilon} \int_{t_0}^t \omega(s) ds\right) \zeta(t),$$

where $\zeta = \pi + i\rho$ solves the differential equation

$$\dot{\zeta}(t) = -\frac{1}{2} \frac{\dot{\omega}(t)}{\omega(t)} \exp\left(-\frac{2i}{\varepsilon} \int_{t_0}^t \omega(s) ds\right) \zeta(t)$$

and satisfies $\zeta(t) = \zeta(t_0)(1 + \mathcal{O}(\varepsilon))$ for $t = \mathcal{O}(1)$. (In the above notation, we have $\eta = \frac{1}{\sqrt{2}}\varepsilon^{-1/2}(\zeta, \bar{\zeta})^T$.) The action

$$I(t) = \frac{1}{\omega(t)} \left(\frac{1}{2} \dot{q}(t)^2 + \frac{\omega(t)^2}{2\varepsilon^2} q(t)^2 \right)$$

is an adiabatic invariant.

XIV.2.2 Adiabatic Integrators

A simple long-time-step integrator for the oscillatory mechanical system with time-dependent Hamiltonian (2.1) now reads as follows:

- Solve the slow limit system (2.29) for p_0, q_0 , e.g., by the Störmer-Verlet method.
- Keep the adiabatic variable η constant at its initial value.

Under the condition of bounded energy (2.1) and the frequency separation condition (2.24), the error in η is then $\mathcal{O}(\varepsilon)$ over intervals $t \leq \text{Const.}$ by (2.26). The difference between the solutions of (2.21) and the limit equation (2.29) is bounded by $\mathcal{O}(\varepsilon^2)$ for $t \leq \text{Const.}$, as can be shown by forming the difference of the equations, integrating, estimating the integral of the extra terms by $\mathcal{O}(\varepsilon^2)$ using (2.26) and partial integration, and applying the Gronwall inequality. In the original variables p, q of (2.1) this yields an error $\mathcal{O}(\varepsilon^2)$ in the positions and $\mathcal{O}(\varepsilon)$ in the momenta.

More refined integrators are needed for two independent reasons:

1. to keep control of η on subintervals where the frequencies are not well separated and where η may thus deviate from its near-constant value;
2. to obtain higher order of approximation on intervals with separated frequencies.

We simplify the following presentation by assuming that the potential U is quadratic:

$$U(q, t) = \frac{1}{2} q^T G(t) q,$$

with a symmetric matrix $G(t)$ depending smoothly on t . We leave the required modifications for general U to the interested reader. Alternatively, the method with $U = 0$ can be used in the splitting approach of Sect. XIV.2.3 below.

An adiabatic integrator as described in Sect. XIV.1.2 can be extended to (2.23) and combined with a symmetric splitting between the weakly coupled systems (2.21) and (2.23): we begin with a symplectic Euler half-step for p_0, q_0 (denoting the time levels by superscripts),

$$\begin{aligned}
p_0^{1/2} &= p_0^0 - \frac{h}{2} \left(L_{00} p_0^{1/2} + (S_{00} + T_0^T G T_0) q_0^0 \right. \\
&\quad \left. + \varepsilon (S_{01} + T_0^T G T_1) \mathcal{Q}_1^- \eta^0 \right) \\
q_0^{1/2} &= q_0^0 + \frac{h}{2} \left(p_0^{1/2} + L_{00}^T q_0^0 + \varepsilon L_{10}^T \mathcal{Q}_1^- \eta^0 \right).
\end{aligned} \tag{2.30}$$

Here the matrix functions L_{00} , L_{10} , S_{00} , S_{01} , T_0 , T_1 are evaluated at $t_{1/2} = t_0 + h/2$, and \mathcal{Q}_1^- is the average of the oscillatory function Q_1 of (2.20) over the half-step,

$$\mathcal{Q}_1^- \approx \frac{2}{h} \int_{t_0}^{t_{1/2}} Q_1(t) dt,$$

obtained with a linear approximation of the phase $\Phi(t)$ and analytic computation of the integral. We then make a full step for η with Eq. (2.23) like in (1.12),

$$\begin{aligned}
\eta^1 &= \eta^0 + h \left(E(\Phi) \bullet \mathcal{I} \bullet W \right) \frac{1}{2} (\eta^1 + \eta^0) \\
&\quad - h \mathcal{P}_1^* \left(L_{10} p_0^{1/2} + (S_{10} + T_1^T G T_0) q_0^{1/2} \right),
\end{aligned} \tag{2.31}$$

where again all matrix functions are evaluated at $t_{1/2}$, and \mathcal{P}_1 is the linear-phase approximation to the average

$$\mathcal{P}_1 \approx \frac{1}{h} \int_{t_0}^{t_1} P_1(t) dt.$$

The matrix W is as in (2.22), but with S_{11} replaced by $S_{11} + T_1^T G T_1$. The step is completed by a half-step for p_0, q_0 with the adjoint symplectic Euler method:

$$\begin{aligned}
p_0^1 &= p_0^{1/2} - \frac{h}{2} \left(L_{00} p_0^{1/2} + (S_{00} + T_0^T G T_0) q_0^1 \right. \\
&\quad \left. + \varepsilon (S_{01} + T_0^T G T_1) \mathcal{Q}_1^+ \eta^1 \right) \\
q_0^1 &= q_0^{1/2} + \frac{h}{2} \left(p_0^{1/2} + L_{00}^T q_0^1 + \varepsilon L_{10}^T \mathcal{Q}_1^+ \eta^1 \right),
\end{aligned} \tag{2.32}$$

where the matrix functions are still evaluated at $t_{1/2}$, and \mathcal{Q}_1^+ approximates the average of Q_1 over the second half-step.

We now give local error bounds for this integrator, under conditions that include the case of an avoided crossing of frequencies.

Theorem 2.2. *Suppose that the functions in (2.1) are smooth and the frequencies satisfy (2.24) with minimal distance $\delta > 0$ for $t_0 \leq t \leq t_0 + h$, and the orthogonal matrix $Q_*(t)$ of (2.7), which diagonalizes the nonsingular part of the stiffness matrix, has derivatives bounded by $\dot{Q}_*(t) = \mathcal{O}(\delta^{-1})$, $\ddot{Q}_*(t) = \mathcal{O}(\delta^{-2})$. Assume further the energy bound (2.2) for the initial values. Then, the local error of method (2.30)–(2.32) is bounded by*

$$\begin{aligned}
p_0^1 - p_0(t_0 + h) &= \mathcal{O}(h^3/\delta^2) + \mathcal{O}(\varepsilon h^2/\delta^2) \\
q_0^1 - q_0(t_0 + h) &= \mathcal{O}(h^3/\delta) + \mathcal{O}(\varepsilon h^2/\delta^2) \\
\eta^1 - \eta(t_0 + h) &= \mathcal{O}(h^2/\delta^2).
\end{aligned}$$

The constants symbolized by \mathcal{O} do not depend on ε , h , and δ .

Proof. (a) Under the given conditions we have

$$\begin{aligned}
K_{00} = \mathcal{O}(1), \quad K_{01} = \mathcal{O}(1), \quad K_{10} = \mathcal{O}(1), \quad K_{11} = \mathcal{O}(\delta^{-1}), \text{ and} \\
\dot{K}_{00} = \mathcal{O}(1), \quad \dot{K}_{01} = \mathcal{O}(\delta^{-1}), \quad \dot{K}_{10} = \mathcal{O}(\delta^{-1}), \quad \dot{K}_{11} = \mathcal{O}(\delta^{-2}),
\end{aligned}$$

This yields the bounds

$$L_{00}, L_{10}, S_{00}, S_{11} = \mathcal{O}(1)$$

and similarly for their derivatives, and

$$L_{11}, S_{01}, S_{10} = \mathcal{O}(\delta^{-1}), \quad \dot{L}_{11}, \dot{S}_{01}, \dot{S}_{10} = \mathcal{O}(\delta^{-2}),$$

and hence also

$$W = \mathcal{O}(\delta^{-1}), \quad \dot{W} = \mathcal{O}(\delta^{-2}).$$

So we have from the energy bound and the differential equation (2.23) for η ,

$$\eta = \mathcal{O}(1), \quad \dot{\eta} = \mathcal{O}(\delta^{-1}).$$

From the differential equations (2.21) for p_0, q_0 we conclude

$$\ddot{p}_0 = \mathcal{O}(\delta^{-1}) + \mathcal{O}(\varepsilon \delta^{-2}), \quad \ddot{q}_0 = \mathcal{O}(\varepsilon \delta^{-1}).$$

(b) To study the local error in η , we integrate (2.23) from t_0 to $t_0 + h$ and compare with the corresponding term in (2.31):

$$\begin{aligned}
&\int_{t_0}^{t_0+h} P_1^*(t) \left(L_{10} p_0 + (S_{10} + T_1^T G T_0) q_0 \right) (t) dt \\
&\quad - h P_1^* \left(L_{10}(t_{1/2}) p_0^{1/2} + (S_{10} + T_1^T G T_0)(t_{1/2}) q_0^{1/2} \right) \\
&= \mathcal{O}(h^2/\delta^2),
\end{aligned}$$

where we have used the above bounds and the error estimate for the linear phase approximation in the average of $P_1(t)$, cf. Sect. XIV.1.2,

$$\mathcal{P}_1 - \frac{1}{h} \int_{t_0}^{t_1} P_1(t) dt = \mathcal{O}(h/\delta).$$

Combining this estimate with the error bound of the adiabatic midpoint rule for the homogeneous equation as given in Theorem 1.2 yields the stated error bound for η_1 .

(c) The error bound for the components p_0, q_0 comes about by combining error bounds for the Störmer–Verlet method (which require the bounds for \ddot{p}_0, \ddot{q}_0) and the estimates

$$\begin{aligned} & \int_{t_0}^{t_0+h/2} \varepsilon(S_{01} + T_0^T G T_1) Q_1 \eta(t) dt - \frac{h}{2} \varepsilon(S_{01} + T_0^T G T_1)(t_{1/2}) Q_1^- \eta^0 \\ &= \mathcal{O}(\varepsilon h^2 / \delta^2) \end{aligned}$$

and

$$\int_{t_0}^{t_0+h/2} \varepsilon L_{10}^T Q_1 \eta(t) dt - \frac{h}{2} \varepsilon L_{10}(t_{1/2}) Q_1^- \eta^0 = \mathcal{O}(\varepsilon h^2 / \delta),$$

and the same estimates for the second half-step. See also Exercise 7 for a similar situation. \square

In the case of well-separated eigenvalues, the global error on bounded time intervals is thus bounded by $\mathcal{O}(h^2) + \mathcal{O}(h\varepsilon)$ in p_0, q_0 for $t \leq \text{Const.}$ and by $\mathcal{O}(h)$ in η . In the original variables p, q of (2.1), this then yields an error

$$q_n - q(t_n) = \mathcal{O}(h^2) + \mathcal{O}(h\varepsilon), \quad p_n - p(t_n) = \mathcal{O}(h) \quad \text{for } t_n \leq \text{Const.}$$

With an adaptive step size strategy as in Sect. XIV.1.2, it is again possible to follow η through non-adiabatic transitions near avoided crossings of eigenvalues.

A higher-order scheme with a global error of $\mathcal{O}(h^2)$ in η – in the situation of separated eigenvalues – is obtained by replacing the upper line in (2.31) by a second-order adiabatic integrator as discussed in Sect. XIV.1.2, leaving the last term in (2.31) unaltered. In the original variables p, q of (2.1), the error is then $\mathcal{O}(h^2)$ both in positions and (fast and slow) momenta. The error is even $\mathcal{O}(\varepsilon h^2)$ in the fast positions q_1 of (2.8), which oscillate with an amplitude $\mathcal{O}(\varepsilon)$. We refer to Lorenz, Jahnke & Lubich (2005) for the particular case of second-order differential equations $\ddot{q} + \varepsilon^{-2} A(t)q = 0$ with a positive definite matrix $A(t)$.

XIV.2.3 Error Analysis of the Impulse Method

The transformation to adiabatic variables of Sect. XIV.2.1 also gives new insight into the error behaviour of multiple time stepping methods such as the impulse or mollified impulse method discussed in Sections VIII.4 and XIII.1, which do not use coordinate transforms in the method formulation. These methods are of interest when the eigendecompositions needed in adiabatic integrators are computationally more expensive than doing many small steps with the fast subsystem, and when evaluations of the potential force are so costly that the computational work for the fast subsystem becomes irrelevant. We consider the splitting

$$H = H^{\text{fast}} + H^{\text{slow}}$$

of the Hamiltonian (2.3) with

$$\begin{aligned}
H^{\text{fast}}(p, E, q, t) &= \frac{1}{2} p^T M(t)^{-1} p + \frac{1}{2\varepsilon^2} q^T A(t) q + E \\
H^{\text{slow}}(p, E, q, t) &= U(q, t).
\end{aligned}$$

The impulse method is given as the composition of the exact flows of the subsystems (see Sections VIII.4 and XIII.1.3):

$$\Phi_h = \varphi_{h/2}^{\text{slow}} \circ \varphi_h^{\text{fast}} \circ \varphi_{h/2}^{\text{slow}},$$

where we are interested in taking long time steps $h \geq c\varepsilon$ (with a positive constant c). The equations of motion of the slow subsystem,

$$\dot{p} = -\nabla U(q, t), \quad \dot{q} = 0, \quad \dot{t} = 0,$$

are solved trivially by

$$\hat{p} = p - \frac{h}{2} \nabla U(q, t), \quad \hat{q} = q, \quad \hat{t} = t.$$

In contrast, the fast subsystem needs to be integrated approximately, e.g., by many small substeps with the Störmer–Verlet method in the original variables (p, q) or by one step of the method (2.30)–(2.32) with $G = 0$ in adiabatic variables (p_0, q_0, η) . In the following we ignore the error resulting from this additional approximation and study the splitting method with exact flows.

The error behaviour of this method can be understood with the help of the transformation to adiabatic variables of Sect. XIV.2.1. The impulse method in the adiabatic variables p_0, q_0, η is obtained by splitting the differential equations (2.21) and (2.23). The fast subsystem is obtained by simply putting $U = 0$ in these equations, and the slow subsystem reads

$$\begin{aligned}
\dot{p}_0 &= -T_0^T \nabla U(T_0 q_0 + \varepsilon T_1 Q_1 \eta, t), \quad \dot{q}_0 = 0 \\
\dot{\eta} &= -P_1^* T_1^T \nabla U(T_0 q_0 + \varepsilon T_1 Q_1 \eta, t)
\end{aligned}$$

along with $\dot{t} = 0$, so that the argument in all the matrices is frozen at the initial time. Here $P_1(t)$ and $Q_1(t)$ are again the highly oscillatory matrix functions of (2.20). Since $Q_1 P_1^* = 0$ we have $Q_1 \eta = \text{Const.}$, and therefore, in these variables the flow $\varphi_{h/2}^{\text{slow}}$ is the mapping given by

$$\begin{aligned}
\hat{p}_0 &= p_0 - \frac{h}{2} T_0^T \nabla U(T_0 q_0 + \varepsilon T_1 Q_1 \eta, t_0), \quad \hat{q}_0 = q_0 \\
\hat{\eta} &= \eta - \frac{h}{2} P_1^* T_1^T \nabla U(T_0 q_0 + \varepsilon T_1 Q_1 \eta, t_0),
\end{aligned} \tag{2.33}$$

where the matrices T_0, T_1, P_1, Q_1 are evaluated at t_0 . In the impulse method, the above values are the starting values for a step with φ_h^{fast} , which is followed by another application of $\varphi_{h/2}^{\text{slow}}$.

A disturbing feature in (2.33) is the appearance of the particular value $P_1(t_0)$ of the highly oscillatory function instead of the average \mathcal{P}_1 as in (2.31).

We now consider the error propagation for η in the case of well-separated frequencies. Recall that the exact solution then satisfies $\eta(t) = \eta(0) + \mathcal{O}(\varepsilon)$ for $t \leq \text{Const.}$ For ease of presentation we consider a constant step size h .

Lemma 2.3. *Assume the energy bound (2.2) for the initial values. If the frequencies $\omega_j(t)$ remain separated from each other, then the result after n steps satisfies, for $nh \leq T \leq \text{Const.}$,*

$$\eta_n = \eta_0 + \sigma_n + \mathcal{O}(\varepsilon), \quad (2.34)$$

where

$$\|\sigma_n\| \leq C\kappa \quad \text{with} \quad \kappa = \max_{0 \leq nh \leq T} \max_k \left\| h \sum_{j=0}^n \exp\left(\frac{i}{\varepsilon} \phi_k(t_j)\right) \right\|. \quad (2.35)$$

Proof. We have $\eta_n = \eta_h(t_n)$, where $\eta_h(t)$ solves the differential equation with impulses,

$$\dot{\eta}_h = \exp\left(-\frac{i}{\varepsilon} \Phi\right) W \exp\left(\frac{i}{\varepsilon} \Phi\right) \eta_h + r + \sum_j \Delta \eta_j \delta_j.$$

Here $W(t)$ is the matrix (2.22) appearing in (2.23), and

$$r(t) = -P_1^*(t) (L_{10}(t)p_{0,h}(t) + S_{01}(t)q_{0,h}(t))$$

with $p_{0,h}(t)$, $q_{0,h}(t)$ denoting the piecewise constant functions that take the values of the numerical solution. Further we have

$$\Delta \eta_j = -h P_1(t_j)^* T_1(t_j)^T \nabla U(T_0(t_j)q_{0,j} + \varepsilon T_1(t_j)Q_1(t_j)\eta_j, t_j),$$

the expression on the right-hand side of (2.33), and δ_j is a Dirac impulse located at t_j . It follows that, for $t = nh$,

$$\begin{aligned} \eta_n - \eta_0 &= \eta_h(t_n) - \eta_h(0) \\ &= \int_0^t \exp\left(-\frac{i}{\varepsilon} \Phi(s)\right) W(s) \exp\left(\frac{i}{\varepsilon} \Phi(s)\right) \eta_h(s) ds + \int_0^t r(s) ds + \sigma_n, \end{aligned}$$

where σ_n is the trapezoidal sum of the terms on the right-hand side of (2.33):

$$\sigma_n = -h \sum_{j=0}' P_1(t_j)^* T_1(t_j)^T \nabla U(T_0(t_j)q_{0,j} + \varepsilon T_1(t_j)Q_1(t_j)\eta_j, t_j). \quad (2.36)$$

The prime on the sum indicates that the first and last term are taken with the factor $\frac{1}{2}$. Using partial integration as in (1.6), we obtain

$$\int_0^t \exp\left(-\frac{i}{\varepsilon} \Phi(s)\right) W(s) \exp\left(\frac{i}{\varepsilon} \Phi(s)\right) \eta_h(s) ds = \mathcal{O}(\varepsilon),$$

and by partial integration as in (2.25),

$$\int_0^t r(s) ds = \mathcal{O}(\varepsilon).$$

This shows (2.34). A partial summation in (2.36), summing up the oscillatory terms $P_1(t_j)$ and differencing the smoother other terms, then yields (2.35). \square

The size of κ of (2.35) depends on possible resonances between the step size and the frequencies, yielding κ between $\mathcal{O}(h)$ and $\mathcal{O}(1)$. For the error of the method we have the following.

Theorem 2.4. *Assume the energy bound (2.2) for the initial values. If the frequencies $\omega_j(t)$ remain separated from each other, then the error of the impulse method after n steps with step size $h \geq c\varepsilon$ satisfies*

$$\begin{aligned} p_n - p(t_n) &= \mathcal{O}(\kappa) \\ q_n - q(t_n) &= \mathcal{O}(h^2) + \mathcal{O}(\varepsilon\kappa). \end{aligned}$$

The constants symbolized by \mathcal{O} do not depend on ε , h and n with $nh \leq \text{Const.}$

Proof. The error of size $\mathcal{O}(\kappa)$ in η immediately implies an error of size $\mathcal{O}(\kappa)$ in the actions $I_j = \frac{1}{2}|\eta_j|^2$, and an error of $\mathcal{O}(\kappa)$ in the fast momenta p_1 and of $\mathcal{O}(\varepsilon\kappa)$ in the fast positions q_1 of (2.9); recall the transformation (2.16) and the rescaling. In the slow components p_0, q_0 the method is a perturbed variant of the Störmer–Verlet method. The contribution of the perturbations $\varepsilon T_1 Q_1 \eta$ to the error is of size $\mathcal{O}(\varepsilon\kappa)$. This is seen by applying the simple lemma below with $y = (p_0, q_0)$ and

$$d_n = \left(-hT_0(t_n)^T \nabla^2 U(T_0(t_n)q_{0,n}, t_n) \varepsilon T_1(t_n)Q_1(t_n)\eta_n \right) + \mathcal{O}(h^2\varepsilon)$$

and using partial summation of the d_n , summing up the oscillatory terms $Q_1(t_n)$ and differencing the other terms. \square

Lemma 2.5. *Let $\Phi_h(y) = y + hF_h(y)$ be a one-step method where F_h has Lipschitz constant L . Consider the method and a perturbation,*

$$y_{n+1} = \Phi_h(y_n) \quad \text{and} \quad \tilde{y}_{n+1} = \Phi_h(\tilde{y}_n) + d_n,$$

with the same starting values $\tilde{y}_0 = y_0$. Then, the difference is bounded by

$$\|\tilde{y}_n - y_n\| \leq e^{nhL} \cdot \max_{0 \leq k \leq n-1} \left\| \sum_{j=0}^k d_j \right\|.$$

Proof. The result follows from

$$\tilde{y}_n - y_n = h \sum_{j=0}^{n-1} (F_h(\tilde{y}_j) - F_h(y_j)) + \sum_{j=0}^{n-1} d_j$$

with the discrete Gronwall inequality. \square

XIV.2.4 Error Analysis of the Mollified Impulse Method

The problem with possible step-size resonances can be greatly alleviated by the mollified impulse method (see Sect. XIII.1.4) where the potential $U(q, t)$ is replaced by a modified potential $\bar{U}(q, t)$. A good choice is

$$\bar{U}(q, t) = U(\mathcal{A}(t)q, t) \quad \text{with} \quad \mathcal{A}(t) = C(t)Q(t) \begin{pmatrix} I & 0 \\ 0 & \mathcal{S}(t) \end{pmatrix} Q(t)^T C(t)^{-1} \quad (2.37)$$

with C and Q of (2.4) and (2.6), and

$$\mathcal{S}(t) = \text{sinc}\left(\frac{h}{\varepsilon} \Omega(t)\right) = \frac{1}{2h} \int_{-h}^h \exp\left(\pm \frac{is}{\varepsilon} \Omega(t)\right) ds.$$

A calculation shows that it replaces (2.33) by

$$\begin{aligned} \hat{p}_0 &= p_0 - \frac{h}{2} T_0^T \nabla U(T_0 q_0 + \varepsilon T_1 \mathcal{Q}_1 \eta, t_0), & \hat{q}_0 &= q_0 \\ \hat{\eta} &= \eta - \frac{h}{2} \mathcal{P}_1^* T_1^T \nabla U(T_0 q_0 + \varepsilon T_1 \mathcal{Q}_1 \eta, t_0), \end{aligned} \quad (2.38)$$

with matrix functions evaluated at t_0 , where $\mathcal{P}_1(t)$ and $\mathcal{Q}_1(t)$ are the linear-phase approximations to the average over the interval $[t-h, t+h]$ of P_1 and Q_1 , respectively,

$$\begin{aligned} \mathcal{P}_1(t) &= \mathcal{S}(t)P_1(t) = \frac{1}{2h} \int_{t-h}^{t+h} P_1(s) ds + \mathcal{O}(h) \\ \mathcal{Q}_1(t) &= \mathcal{S}(t)Q_1(t) = \frac{1}{2h} \int_{t-h}^{t+h} Q_1(s) ds + \mathcal{O}(h). \end{aligned}$$

Therefore, (2.34) and (2.36) hold with the highly oscillatory $P_1(t_j)$ replaced by the averages $\mathcal{P}_1(t_j)$. Using a partial summation in (2.36) and noting that, for $t = nh \leq \text{Const.}$,

$$\left\| h \sum_{j=1}^n \mathcal{P}_1(t_j) \right\| = \left\| \int_0^t P_1(s) ds \right\| + \mathcal{O}(h) = \mathcal{O}(\varepsilon) + \mathcal{O}(h),$$

we obtain an estimate

$$\eta_n = \eta_0 + \mathcal{O}(h)$$

instead of the corresponding bound (2.34) with (2.35). This eliminates the bad effect of step size resonances (large κ) on the propagation in the fast variables over bounded time intervals $t \leq \text{Const.}$ (though not on longer intervals, as we know from Chap. XIII). The more harmless effect of step size resonances on the slow variables, as visible in the term $\mathcal{O}(\varepsilon\kappa)$ in Theorem 2.4, is likewise reduced to $\mathcal{O}(\varepsilon h)$. We thus obtain the following improvement over the error bounds in Theorem 2.4.

Theorem 2.6. *Assume the energy bound (2.2) for the initial values. If the frequencies $\omega_j(t)$ remain separated from each other, then the error of the above mollified impulse method after n steps with step size $h \geq c\varepsilon$ satisfies*

$$\begin{aligned} p_n - p(t_n) &= \mathcal{O}(h) \\ q_n - q(t_n) &= \mathcal{O}(h^2). \end{aligned}$$

The constants symbolized by \mathcal{O} do not depend on ε , h and n with $nh \leq \text{Const.}$ \square

A direct implementation of this method requires just the same matrix decompositions that are needed for the integrators in adiabatic variables. It is then reasonable to use one step of the adiabatic integrator of Sect. XIV.2.2 for solving the fast subsystem over a time step.

An alternative is to compute the average $\mathcal{A}(t)$ by small time steps from the linear differential equation with the Hamiltonian H^{fast} , as formulated in Sect. XIII.1.4. The method described here then corresponds to (XIII.1.18) with $c = 1$.

XIV.3 Mechanical Systems with Solution-Dependent Frequencies

We² consider the Hamiltonian

$$H(p, q) = \frac{1}{2} p^T M(q)^{-1} p + U(q) + \frac{1}{\varepsilon^2} V(q) \quad (3.1)$$

with a strong potential $\varepsilon^{-2}V(q)$ that penalizes some directions of motion. Analytical studies of this problem were done by Rubin & Ungar (1957), Takens (1980), and Bornemann (1998). In an alternative approach to these works, we here describe a transformation of the problem to adiabatic variables. This gives new insight into the solution behaviour and can be used as the starting point for the construction of long-time-step integrators. It also enables us to analyse the error of multiple time-stepping methods.

XIV.3.1 Constraining Potentials

We consider the Hamiltonian (3.1), where $M(q)$ is a symmetric positive definite mass matrix depending smoothly on the positions $q \in \mathbb{R}^n$, U is a smooth potential, and the constraining potential is assumed to satisfy the following:

² This section was written in cooperation with Katina Lorenz (Doctoral Thesis, Univ. Tübingen, in preparation).

The smooth function $V : D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ attains its minimum value 0 on a d -dimensional manifold $\mathcal{V} \subset \mathbb{R}^n$,

$$\mathcal{V} = \{q \in D \mid V(q) = \min V = 0\}. \quad (3.2)$$

In a neighbourhood of \mathcal{V} , the potential V is strongly convex along directions non-tangential to \mathcal{V} , that is, there exists $\alpha > 0$ such that for $q \in \mathcal{V}$, the Hessian $\nabla^2 V(q)$ satisfies

$$v^T \nabla^2 V(q) v \geq \alpha \cdot v^T M(q) v \quad (3.3)$$

for all vectors v in the $M(q)$ -orthogonal complement of the tangent space $T_q \mathcal{V}$.

We let $m = n - d$ be the number of independent constraints that locally describe the manifold \mathcal{V} .

Example 3.1 (Chain of Stiff Springs). The position of $m + 1$ mass points in a plane, arranged in a chain connected by stiff springs with spring constants α_i^2/ε^2 , is determined by the Cartesian coordinates of the first mass point and by m angles φ_i and the elongations d_i of the m springs. The constraining potential is

$$V = \frac{1}{2} \sum_{i=1}^m \alpha_i^2 d_i^2,$$

and the constraint manifold is described by $d_1 = \dots = d_m = 0$ corresponding to non-elongated springs. The frequencies of the vibrations in such a chain depend on the angles.

In the above example we have, in the coordinates given by the angles and elongations, a potential V of the form

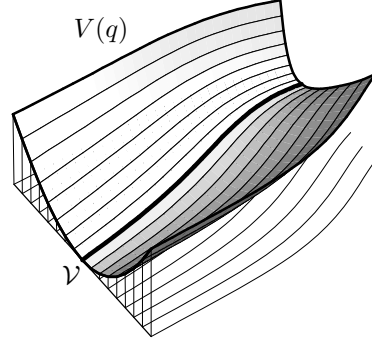
$$V(q) = \frac{1}{2} q_1^T A(q_0) q_1 \quad (3.4)$$

for $q = (q_0, q_1) \in \mathbb{R}^d \times \mathbb{R}^m$, with a positive definite matrix $A(q_0)$. The manifold of constraints is here simply $\mathcal{V} = \mathbb{R}^d \times 0$. As the following lemma shows, this is already the general situation in suitable local coordinates.

Lemma 3.2. *Under conditions (3.2)–(3.3), there exists a smooth local change of coordinates $q = \chi(y)$ such that*

$$V(q) = \frac{1}{2} y_1^T A(y_0) y_1 \quad \text{for } q = \chi(y)$$

with $y = (y_0, y_1)$ near 0 in $\mathbb{R}^d \times \mathbb{R}^m$, where $A(y_0)$ is a symmetric positive definite $m \times m$ matrix.



Proof. In a first step, we choose local coordinates $q = \psi(x)$ with $x = (x_0, x_1)$ near 0 in $\mathbb{R}^d \times \mathbb{R}^m$, such that $q = \psi(x) \in \mathcal{V}$ if and only if $x_1 = 0$. In these coordinates, denoting $\widehat{V}(x) = V(q)$ for $q = \psi(x)$, we then have

$$\widehat{V}(x_0, 0) = 0, \quad \nabla \widehat{V}(x_0, 0) = 0$$

by (3.2), and

$$A(x_0) := \nabla_{x_1}^2 \widehat{V}(x_0, 0) \quad \text{is positive definite}$$

by (3.3). We now change coordinates by the near-identity transformation

$$y_0 = x_0, \quad y_1 = \mu(x)x_1$$

where the real factor $\mu(x)$ (near 1 for x_1 near 0) is to be chosen such that

$$\frac{1}{2} y_1^T A(y_0) y_1 = \widehat{V}(x_0, x_1).$$

Since the right-hand side equals

$$\widehat{V}(x_0, x_1) - \widehat{V}(x_0, 0) - x_1^T \nabla \widehat{V}(x_0, 0) = \frac{1}{2} x_1^T A(x_0) x_1 + r(x)$$

with $r(x) = \mathcal{O}(\|x_1\|^3)$, the choice

$$\mu(x) = \sqrt{1 + \frac{2r(x)}{x_1^T A(x_0) x_1}}$$

does the trick. \square

We remark that Lemma 3.2 could be obtained as a corollary to the Morse lemma, for which we refer to Abraham & Marsden (1978) and Crouzeix & Rappaz (1989).

The change to the local coordinates $x = (x_0, x_1)$ such that $V(q) = 0$ if and only if $x_1 = 0$ for $q = \psi(x)$, is not numerically constructive from the mere knowledge of an expression for the potential V . However, in many situations the manifold \mathcal{V} can be described by constraints $g(q) = 0$, and $x_1 = g$ can then be extended to a full set of coordinates. The above transformation from x to y can be done numerically. In the usual way, the transformation $q = \chi(y)$ of the position coordinates extends to a canonical transformation by setting $p_y = \chi'(y)^T p$ for the conjugate momenta; see Example VI.5.2.

Solutions of (3.1) are in general oscillatory with frequencies of size $\sim \varepsilon^{-1}$. There exist, however, special solutions having arbitrarily many time derivatives bounded independently of ε , which for arbitrary $N \geq 1$ stay $\mathcal{O}(\varepsilon^N)$ close to a manifold $\mathcal{V}^{\varepsilon, N}$ that has a distance $\mathcal{O}(\varepsilon)$ to \mathcal{V} . See Lubich (1993), where also implicit Runge-Kutta methods for the approximation of the smooth solutions are studied. In this section we are, however, interested in approximating general oscillatory solutions of bounded energy.

XIV.3.2 Transformation to Adiabatic Variables

We start from a Hamiltonian (3.1) in coordinates (p, q) where the constraining potential is already of the form (3.4) for $q = (q_0, q_1)$. We note that for a system of bounded energy, we then have $q_1 = \mathcal{O}(\varepsilon)$.

We now perform a series of canonical transformations that take the Hamiltonian into a form that is better suited for a direct numerical treatment and for the error analysis of multiple time-stepping methods. The transformations are similar to those for the time-dependent case treated in Sect. XIV.2.1, but here they appear in a permuted order.

Transforming the Stiffness Matrix into the Identity. We write the Cholesky decomposition of the stiffness matrix as

$$A(q_0) = C(q_0)^{-T} C(q_0)^{-1}$$

and change to variables

$$q_0 = \tilde{q}_0, \quad q_1 = C(\tilde{q}_0) \tilde{q}_1$$

along with the conjugate momenta

$$\tilde{p}_0 = p_0 + \left(\frac{\partial}{\partial \tilde{q}_0} C(\tilde{q}_0) \tilde{q}_1 \right)^T p_1, \quad \tilde{p}_1 = C(\tilde{q}_0)^T p_1.$$

With the transformed mass matrix $\tilde{M}(\tilde{q}) = B(\tilde{q}) M(\tilde{q}_0, C(\tilde{q}_0) \tilde{q}_1) B(\tilde{q})^T$ (for the matrix $B(\tilde{q})$ that transforms $\tilde{p} = B(\tilde{q}) p$) and the potential $\tilde{U}(\tilde{q}) = U(\tilde{q}_0, C(\tilde{q}_0) \tilde{q}_1)$, the Hamiltonian takes the simplified form (we omit all tildes)

$$H = \frac{1}{2} p^T M(q)^{-1} p + \frac{1}{2\varepsilon^2} q_1^T q_1 + U(q). \quad (3.5)$$

Eliminating Off-Diagonal Blocks in the Mass Matrix. We write the mass matrix $M(q)$ as

$$M = \begin{pmatrix} M_{00} & M_{01} \\ M_{10} & M_{11} \end{pmatrix}.$$

With $G(\bar{q}_0) = -M_{00}(\bar{q}_0, 0)^{-1} M_{01}(\bar{q}_0, 0)$, we transform

$$q_0 = \bar{q}_0 + G(\bar{q}_0) \bar{q}_1, \quad q_1 = \bar{q}_1,$$

with the conjugate momenta

$$\bar{p}_0 = p_0 + \left(\frac{\partial}{\partial \bar{q}_0} G(\bar{q}_0) \bar{q}_1 \right)^T p_0, \quad \bar{p}_1 = p_1 + G(\bar{q}_0)^T p_0.$$

This canonical change of variables eliminates M_{01} and M_{10} in the transformed mass matrix $\bar{M}(q_0, 0)$ and keeps the Schur complement on the block diagonal: with the symmetric positive definite matrices

$$\overline{M}_0(\overline{q}_0) = M_{00}(\overline{q}_0, 0), \quad \overline{M}_1(\overline{q}_0) = (M_{11} - M_{10}M_{00}^{-1}M_{01})(\overline{q}_0, 0),$$

the transformation puts the Hamiltonian into the form (we omit all bars)

$$\begin{aligned} H = & \frac{1}{2} p_0^T M_0(q_0)^{-1} p_0 + \frac{1}{2} p_1^T M_1(q_0)^{-1} p_1 + \frac{1}{2\varepsilon^2} q_1^T q_1 \\ & + \frac{1}{2} p^T R(q) p + U(q_0 + G(q_0)q_1, q_1) \end{aligned} \quad (3.6)$$

where R is a smooth matrix-valued function satisfying

$$R(q_0, 0) = 0. \quad (3.7)$$

Diagonalizing the Mass Matrix of the Fast Variables. We diagonalize

$$M_1(q_0) = Q(q_0)\Omega(q_0)^{-2}Q(q_0)^T$$

with the diagonal matrix $\Omega(q_0) = \text{diag}(\omega_j(q_0))$ of frequencies and an orthogonal matrix $Q(q_0)$, which depends smoothly on q_0 if the frequencies are separated. We transform

$$q_0 = \widehat{q}_0, \quad q_1 = Q(\widehat{q}_0)\widehat{q}_1$$

with the conjugate momenta

$$\widehat{p}_0 = p_0 + \left(\frac{\partial}{\partial \widehat{q}_0} Q(\widehat{q}_0)\widehat{q}_1 \right)^T p_1, \quad \widehat{p}_1 = Q(\widehat{q}_0)^T p_1.$$

The matrix

$$Y(\widehat{q}) = \left(\frac{\partial}{\partial \widehat{q}_0} Q(\widehat{q}_0)\widehat{q}_1 \right)^T Q(\widehat{q}_0)$$

is of size $\mathcal{O}(\widehat{q}_1)$ but it is this expression which may become large near avoided crossings of eigenvalues. We consider the associated matrix

$$X(\widehat{q}) = \begin{pmatrix} 0 & X_{01} \\ X_{10} & X_{11} \end{pmatrix} = \begin{pmatrix} 0 & -M_0^{-1}Y \\ -Y^T M_0^{-1} & Y^T M_0^{-1}Y \end{pmatrix}. \quad (3.8)$$

With a matrix $\widehat{R}(\widehat{q})$ satisfying (3.7), which is a sum of the appropriately transformed previous matrix R and the above matrix X , the Hamiltonian in the new variables $(\widehat{p}, \widehat{q})$ becomes (we omit all hats)

$$\begin{aligned} H = & \frac{1}{2} p_0^T M_0(q_0)^{-1} p_0 + \frac{1}{2} p_1^T \Omega(q_0)^2 p_1 + \frac{1}{2\varepsilon^2} q_1^T q_1 \\ & + \frac{1}{2} p^T R(q) p + U(q_0 + GQ(q_0)q_1, Q(q_0)q_1). \end{aligned} \quad (3.9)$$

Rescaling Positions and Momenta. We change to rescaled fast variables

$$q_0 = \check{q}_0, \quad q_1 = \varepsilon^{1/2} \Omega(\check{q}_0)^{1/2} \check{q}_1$$

(note that $q_1 = \mathcal{O}(\varepsilon)$ implies $\check{q}_1 = \mathcal{O}(\varepsilon^{1/2})$) with the conjugate momenta

$$\check{p}_0 = p_0 + \varepsilon^{1/2} \left(\frac{\partial}{\partial \check{q}_0} \Omega(\check{q}_0)^{1/2} \check{q}_1 \right)^T p_1, \quad \check{p}_1 = \varepsilon^{1/2} \Omega(\check{q}_0)^{1/2} p_1.$$

In the new variables, the Hamiltonian becomes (we omit the hačeks on all variables)

$$\begin{aligned} H &= \frac{1}{2} p_0^T M_0(q_0)^{-1} p_0 + \frac{1}{2\varepsilon} p_1^T \Omega(q_0) p_1 + \frac{1}{2\varepsilon} q_1^T \Omega(q_0) q_1 \\ &\quad + \frac{1}{2} p^T R(q) p + U(T(q_0)q), \end{aligned} \quad (3.10)$$

where

$$T = \left(T_0 \mid \varepsilon^{1/2} T_1 \right) = \begin{pmatrix} I & \varepsilon^{1/2} G Q \Omega^{1/2} \\ 0 & \varepsilon^{1/2} Q \Omega^{1/2} \end{pmatrix}$$

and $R(q)$ is a symmetric matrix of the form

$$R(q) = \begin{pmatrix} R_{00}(q_0, \varepsilon^{1/2} q_1) & \varepsilon^{-1/2} R_{01}(q_0, \varepsilon^{1/2} q_1) \\ \varepsilon^{-1/2} R_{10}(q_0, \varepsilon^{1/2} q_1) & \varepsilon^{-1} R_{11}(q_0, \varepsilon^{1/2} q_1) \end{pmatrix}$$

with smooth functions R_{ij} satisfying $R_{ij}(q_0, 0) = 0$. Therefore, the expression $\frac{1}{2} p^T R(q) p$ can be rewritten in the form

$$\begin{aligned} \frac{1}{2} p^T R(q) p &= \varepsilon^{1/2} c(p_0, q_0)^T q_1 + p_1^T L(p_0, q_0)^T q_1 \\ &\quad + \varepsilon^{-1/2} \tau(p_1, p_1, q_1; p_0, q_0) + \rho(p, q), \end{aligned} \quad (3.11)$$

with a vector c , a matrix L , a function τ that is trilinear in p_1, p_1, q_1 , and a remainder of size $\rho(p, q) = \mathcal{O}(\varepsilon^2)$ for $p_1, q_1 = \mathcal{O}(\varepsilon^{1/2})$, whose partial derivatives with respect to p_1, q_1 are of size $\mathcal{O}(\varepsilon^{3/2})$, and with respect to p_0, q_0 of size $\mathcal{O}(\varepsilon^2)$.

Equations of Motion. The differential equations now take the form

$$\begin{aligned} \dot{p}_0 &= -\nabla_{q_0} \left(\frac{1}{2} p_0^T M_0(q_0)^{-1} p_0 + U(q_0, 0) \right) \\ &\quad - \nabla_{q_0} \left(\frac{1}{2\varepsilon} p_1^T \Omega(q_0) p_1 + \frac{1}{2\varepsilon} q_1^T \Omega(q_0) q_1 \right) + f_0(p, q) \\ \dot{q}_0 &= M_0(q_0)^{-1} p_0 + g_0(p, q) \\ \begin{pmatrix} \dot{p}_1 \\ \dot{q}_1 \end{pmatrix} &= \frac{1}{\varepsilon} \begin{pmatrix} 0 & -\Omega(q_0) \\ \Omega(q_0) & 0 \end{pmatrix} \begin{pmatrix} p_1 \\ q_1 \end{pmatrix} + \begin{pmatrix} f_1(p, q) \\ g_1(p, q) \end{pmatrix} \end{aligned} \quad (3.12)$$

with the functions

$$\begin{pmatrix} f_0 \\ f_1 \end{pmatrix} = -\nabla_q \left(\frac{1}{2} p^T R(q) p + U(T(q_0)q) - U(q_0, 0) \right)$$

$$\begin{pmatrix} g_0 \\ g_1 \end{pmatrix} = R(q)p.$$

We note the magnitudes $f_0 = \mathcal{O}(\varepsilon)$, $g_0 = \mathcal{O}(\varepsilon)$ and $f_1 = \mathcal{O}(\varepsilon^{1/2})$, $g_1 = \mathcal{O}(\varepsilon^{1/2})$ in the case of separated eigenfrequencies, where the diagonalization is smooth with bounded derivatives. By (3.11) we have (omitting the arguments p_0, q_0 in c, L, T)

$$\begin{aligned} f_1 &= -\varepsilon^{1/2}c - Lp_1 + \varepsilon^{-1/2}a(p_1, p_1; p_0, q_0) - \varepsilon^{1/2}T_1^T \nabla U(q_0, 0) + \mathcal{O}(\varepsilon^{3/2}) \\ g_1 &= L^T q_1 + \varepsilon^{-1/2}b(p_1, q_1; p_0, q_0) + \mathcal{O}(\varepsilon^{3/2}) \end{aligned} \quad (3.13)$$

where the functions a and b are bilinear in their first two arguments.

The System in Adiabatic Variables. We finally leave the canonical framework and transform to adiabatic variables as in (2.16). Along a solution $(p(t), q(t))$ of the system (3.12) we consider the diagonal phase matrix $\Phi(t)$ defined by

$$\dot{\Phi} = \Lambda(q_0) \quad \text{with} \quad \Lambda(q_0) = \begin{pmatrix} \Omega(q_0) & 0 \\ 0 & -\Omega(q_0) \end{pmatrix}.$$

With the constant unitary matrix Γ of (2.14), which diagonalizes the matrix in (3.12), we introduce the adiabatic variables

$$\eta = \varepsilon^{-1/2} \exp\left(-\frac{i}{\varepsilon}\Phi\right) \Gamma^* \begin{pmatrix} p_1 \\ q_1 \end{pmatrix} \quad (3.14)$$

and denote the inverse transform as

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \varepsilon^{1/2} \begin{pmatrix} P_1 \\ Q_1 \end{pmatrix} \eta = \varepsilon^{1/2} \Gamma \exp\left(\frac{i}{\varepsilon}\Phi\right) \eta. \quad (3.15)$$

The differential equations (3.12) for p_1, q_1 then turn into

$$\dot{\eta} = \varepsilon^{-1/2} \exp\left(-\frac{i}{\varepsilon}\Phi\right) \Gamma^* \begin{pmatrix} f_1 \\ g_1 \end{pmatrix} = \varepsilon^{-1/2} P_1^* f_1 + \varepsilon^{-1/2} Q_1^* g_1$$

with the arguments $(p_0, \varepsilon^{1/2}P_1\eta, q_0, \varepsilon^{1/2}Q_1\eta)$ in the functions f_1, g_1 . Inserting the expressions for f_1 and g_1 from (3.13), we obtain as in (2.22) and (2.23), with

$$W = -\frac{1}{2} \begin{pmatrix} L - L^T & L + L^T \\ L + L^T & L - L^T \end{pmatrix}, \quad (3.16)$$

the differential equation

$$\dot{\eta} = \exp\left(-\frac{i}{\varepsilon}\Phi\right) W(p_0, q_0) \exp\left(\frac{i}{\varepsilon}\Phi\right) \eta \quad (3.17)$$

$$+ \exp\left(-\frac{i}{\varepsilon}\Phi\right) \Gamma^* \begin{pmatrix} a(P_1\eta, P_1\eta; p_0, q_0) \\ b(P_1\eta, Q_1\eta; p_0, q_0) \end{pmatrix} \quad (3.18)$$

$$- P_1^* \left(c(p_0, q_0) + T_1(q_0)^T \nabla U(q_0, 0) \right) + r \quad (3.19)$$

with the remainder $r(p_0, q_0, P_1\eta, Q_1\eta) = \mathcal{O}(\varepsilon)$.

Adiabatic Invariants. For a solution with bounded energy, both $p_1(t)$ and $q_1(t)$ in (3.12) are of size $\mathcal{O}(\varepsilon^{1/2})$ and hence

$$\eta(t) = \mathcal{O}(1).$$

We now integrate both sides of the above differential equation from 0 to t . The integral of the terms in (3.19) is $\mathcal{O}(\varepsilon)$, as is seen by partial integration since $P_1^*(t)$ is oscillatory with an $\mathcal{O}(\varepsilon)$ integral and p_0, q_0 have bounded derivatives.

We now suppose that the eigenfrequencies $\omega_j(t) := \omega_j(q_0(t))$ remain separated and bounded away from 0: there is a constant $\delta > 0$ such that for any pair $\omega_j(t)$ and $\omega_k(t)$ with $j \neq k$, the lower bounds

$$|\omega_j(t) - \omega_k(t)| \geq \delta, \quad \omega_j(t) \geq \frac{\delta}{2} \quad (3.20)$$

hold for all t under consideration. In this situation, as in Sect. XIV.2.1, the integral from 0 to t of the term (3.17) is bounded by $\mathcal{O}(\varepsilon)$, since the matrix W has zero diagonal.

It remains to study the term (3.18) with the bilinear functions a and b . This term has only oscillatory components if the following non-resonance condition is satisfied: for all j, k, l and all combinations of signs,

$$|\omega_j(t) \pm \omega_k(t) \pm \omega_l(t)| \geq \delta \quad (3.21)$$

with a positive δ independent of ε . In this case, also the integral over the term (3.18) is of size $\mathcal{O}(\varepsilon)$, and we obtain

$$\eta(t) = \eta(0) + \mathcal{O}(\varepsilon) \quad \text{for } t \leq \text{Const.} \quad (3.22)$$

If condition (3.21) is weakened to requiring that for all $j, k, l = 1, \dots, m$,

$$\omega_j(t) \pm \omega_k(t) \pm \omega_l(t) \text{ has a finite number of at most simple zeros} \quad (3.23)$$

in the considered time interval, then the estimate deteriorates to (see Exercise 1)

$$\eta(t) = \eta(0) + \mathcal{O}(\varepsilon^{1/2}) \quad \text{for } t \leq \text{Const.} \quad (3.24)$$

The actions

$$I_j = |\eta_j|^2 \quad (j = 1, \dots, m) \quad (3.25)$$

are thus adiabatic invariants:

$$I_j(t) = I_j(0) + \mathcal{O}(\varepsilon) \quad \text{for } t \leq \text{Const.} \quad (3.26)$$

in case of (3.22), and up to $\mathcal{O}(\varepsilon^{1/2})$ in case of (3.24).

The Slow System. Since the oscillatory energy equals

$$\frac{1}{2\varepsilon} p_1^T \Omega(q_0) p_1 + \frac{1}{2\varepsilon} q_1^T \Omega(q_0) q_1 = \sum_{j=1}^m I_j \omega_j(q_0),$$

the differential equations (3.12) for the slow variables p_0, q_0 become, up to $\mathcal{O}(\varepsilon)$,

$$\begin{aligned} \dot{p}_0 &= -\nabla_{q_0} \left(\frac{1}{2} p_0^T M_0(q_0)^{-1} p_0 + U(q_0, 0) \right) - \sum_{j=1}^m I_j \nabla_{q_0} \omega_j(q_0) \\ \dot{q}_0 &= M_0(q_0)^{-1} p_0. \end{aligned} \quad (3.27)$$

Compared with the constrained system with Hamiltonian $\frac{1}{2} p^T M(q)^{-1} p + U(q)$ on the configuration manifold \mathcal{V} , the slow motion is thus driven by the additional potential $\sum_{j=1}^m I_j \omega_j(q_0)$ depending on the actions I_j . See also Rubin & Ungar (1957), Takens (1980), and Bornemann (1998) for different derivations and discussions of the correction potential.

Avoided Crossing of Frequencies and Takens Chaos. If the distance δ of frequencies in (3.20) becomes so small at a point $q_0(t)$ that $\delta^2 \leq \varepsilon$, then there can again occur $\mathcal{O}(1)$ changes in adiabatic invariants I_j , as in the Zener example of Sect. XIV.1.1. In the present situation of solution-dependent frequencies, however, the level to which I_j jumps after the avoided crossing, depends very sensitively on the slow solution variables $q_0(t)$ through the terms $\exp(\pm \frac{i}{\varepsilon} \Phi)$ in (3.17). In turn, the slow motion of p_0, q_0 after the avoided crossing depends on the new values of I_j through (3.27). The effect is that the slow motion depends very sensitively on perturbations of the initial values in the case of an avoided crossing; see Takens (1980). The indeterminacy of the slow motion in the limit $\varepsilon \rightarrow 0$ is termed *Takens chaos* by Bornemann (1998).

XIV.3.3 Integrators in Adiabatic Variables

A long-time-step integrator for the oscillatory mechanical system with Hamiltonian (3.1) can now be obtained as follows:

Solve the slow system (3.27) in tandem with applying an adiabatic integrator (see Sect. XIV.1.2) to a simplified equation for the adiabatic variables,

$$\dot{\eta} = \exp\left(-\frac{i}{\varepsilon} \Phi\right) W \exp\left(\frac{i}{\varepsilon} \Phi\right) \eta,$$

where W is given by (3.16) with a simplified matrix L : with $v_0 = M_0(q_0)^{-1} p_0$, let

$$L(p_0, q_0) = -\Omega(q_0)^{1/2} \frac{d}{d\tau} \Big|_{\tau=0} Q(q_0 + \tau v_0)^T Q(q_0) \Omega(q_0)^{-1/2}.$$

This matrix L captures the principal terms, coming from the matrix X_{01} in (3.8), which are responsible for a change of the adiabatic invariants due to an avoided

crossing as long as the frequency separation condition (3.20) holds with a possibly ε -dependent $\delta \gg \varepsilon$, e.g., with $\delta \sim \varepsilon^{1/2}$ where $O(1)$ changes occur in the adiabatic invariants. Because of the Takens chaos, it cannot be expected that such an integrator yields a good approximation to “the” solution, but the method can approximate an almost-solution (having a small defect in the differential equations) that passes through the avoided crossing zone, and it detects the change of adiabatic invariants. The properties of integrators of this type are currently under investigation (Lorenz & Lubich 2006).

Further we refer to Jahnke (2003, 2004b) for the construction and analysis of adiabatic integrators for mixed quantum-classical molecular dynamics, where similarly a nonlinear coupling of slow and fast, oscillatory motions occurs.

XIV.3.4 Analysis of Multiple Time-Stepping Methods

The error behaviour of the impulse and mollified impulse method applied to an oscillatory Hamiltonian system (3.1) with well-separated frequencies can be analysed in the adiabatic variables in the same way as we did in Sections XIV.2.3 and XIV.2.4 for the case of time-dependent frequencies. Analogous formulas and the same conclusions hold; essentially we need to replace the argument t by q_0 in the appearing functions. However, their behaviour in the situation of an avoided crossing with Takens chaos is presently not understood.

XIV.4 Exercises

1. Show that

$$\int_0^t \exp\left(\frac{i}{\varepsilon} \phi(s)\right) ds = \mathcal{O}(\varepsilon^{1/(m+1)})$$

if $\lambda := \dot{\phi}$ has finitely many zeros of order at most m in the interval $[0, t]$.

Hint: Use the *method of stationary phase*; see, e.g., Olver (1974) or van der Corput (1934).

2. Show that the adiabatic variables $\eta(t)$ of (1.4) remain approximately constant also in the following cases of non-separated eigenvalues:
 - (a) a multiple eigenvalue $\lambda_j(t)$ of constant multiplicity m for all t and the orthogonal basis $v_{j,1}(t), \dots, v_{j,m}(t)$ of the corresponding eigenspace chosen such that the derivatives $\dot{v}_{j,l}(t)$ are orthogonal to the eigenspace for all t ;
 - (b) a crossing of eigenvalues, $\lambda_j(t_*) = \lambda_k(t_*)$ with $\dot{\lambda}_j(t_*) \neq \dot{\lambda}_k(t_*)$, for which the eigenvectors are smooth functions of t in a neighbourhood of t_* ; see also Born & Fock (1928) for crossings where $\lambda_j - \lambda_k$ can have zeros of higher multiplicity.
3. Let the differential equation (1.1) with smooth skew-hermitian $Z(t)$ be transformed locally over $[t_0, t_0 + h]$ to $z(t) = \exp(-\frac{t}{\varepsilon} Z_*)y(t)$, so that

$$\dot{z} = \frac{1}{\varepsilon} \exp\left(-\frac{t}{\varepsilon} Z_*\right) (Z(t) - Z_*) \exp\left(\frac{t}{\varepsilon} Z_*\right) z$$

with $Z_* = Z(t_0 + h/2)$. Consider the averaged midpoint rule

$$z_1 = z_0 + \frac{1}{\varepsilon} \int_0^h \exp\left(-\frac{s}{\varepsilon} Z_*\right) (\tilde{Z}(s) - Z_*) \exp\left(\frac{s}{\varepsilon} Z_*\right) ds \frac{1}{2}(z_0 + z_1), \quad (4.1)$$

where $\tilde{Z}(t)$ is the quadratic interpolation polynomial through $Z(t_0)$, Z_* , $Z(t_1)$. Show that the local error $z_1 - z(t_1)$ is of size $\mathcal{O}(h^4/\varepsilon^2)$, which is $\mathcal{O}(h^2)$ only for $h = \mathcal{O}(\varepsilon)$. Explain why the error bound cannot be improved to $\mathcal{O}(h^2)$ for $h = \mathcal{O}(\varepsilon^\alpha)$ with $\alpha < 1$.

Hint: See the proofs of Theorems 2.1(i) and 3.1 in Hochbruck & Lubich (1999b), cf. also Iserles (2004).

4. In the situation of the previous exercise, let U be a unitary matrix of eigenvectors of Z_* , and let $\tilde{D}(t)$ be the diagonal matrix containing the diagonal entries of $U^*(\tilde{Z}(t) - Z_*)U$. Find a modification of the above averaged midpoint rule by terms that use only $\tilde{D}(t)$, such that the local error is $\mathcal{O}(h^2)$ for $h \leq \varepsilon^{3/4}$ if the eigenvalues of Z_* are all separated by a distance δ independent of ε .
5. Compare the error behaviour of the averaged midpoint rules (1.12) and (4.1) near the avoided crossing of the eigenvalues in the Zener matrix (1.9).
6. Formulate symmetric modifications of the adiabatic integrators (1.12) and (1.13) that use function evaluations at the grid points t_n and t_{n+1} instead of $t_{n+1/2}$.
7. Consider the differential equation $\dot{y} = f(y) + g(t)$ with a smooth function $f(y)$ and a function $g(t) = \mathcal{O}(1)$ with $\dot{g}(t) = \mathcal{O}(\delta^{-1})$ with respect to a small parameter δ . For the modified midpoint rule

$$y_1 = y_0 + hf\left(\frac{y_0 + y_1}{2}\right) + \int_{t_0}^{t_1} g(t) dt,$$

show that the local error satisfies $y_1 - y(t_1) = \mathcal{O}(h^3/\delta)$.

8. Write the Hamiltonian system (XIII.9.2) in adiabatic variables and relate this to the first terms of the modulated Fourier expansion.
9. Compare the impulse method of Sect. XIV.2.3 with the method based on the splitting

$$H = \left(\frac{1}{2} p^T M(t)^{-1} p + \frac{1}{2\varepsilon^2} q^T A(t) q\right) + (U(q, t) + E).$$

10. Show that Theorem 2.6 remains valid for the choice $\mathcal{S}(t) = 0$ in (2.37). This corresponds to the projection to the constraint manifold in the mollified impulse method as proposed by Izaguirre, Reich & Skeel (1999).

Chapter XV.

Dynamics of Multistep Methods

Multistep methods are the basis of important codes for nonstiff differential equations (Adams methods) and for stiff problems (BDF methods). We study here their applicability to long-time integrations of Hamiltonian or reversible systems.

This chapter starts with numerical experiments which illustrate that the long-time behaviour of classical multistep methods is in general disappointing. They either behave as non-symplectic and non-symmetric one-step methods, or they exhibit undesired instabilities (parasitic solutions). Certain multistep methods for second order equations or partitioned multistep methods, however, have a much better long-time behaviour. They are promising methods, because in a constant step size mode they can be easily implemented, and high order can be obtained with one function evaluation per step. We characterize such methods by studying their underlying one-step method, their symplecticity, their conservation properties, as well as their long-term stability.

XV.1 Numerical Methods and Experiments

We present the numerical methods treated in this chapter, and in numerical experiments we look at their behaviour on Hamiltonian systems.

XV.1.1 Linear Multistep Methods

For first order systems of differential equations $\dot{y} = f(y)$, linear multistep methods are defined by the formula

$$\sum_{j=0}^k \alpha_j y_{n+j} = h \sum_{j=0}^k \beta_j f(y_{n+j}), \quad (1.1)$$

where α_j, β_j are real parameters, $\alpha_k \neq 0$, and $|\alpha_0| + |\beta_0| > 0$. For an application of this formula we need a starting procedure which, in addition to an initial value $y(t_0) = y_0$, provides approximations y_1, \dots, y_{k-1} to $y(t_0+h), \dots, y(t_0+(k-1)h)$. The approximations y_n to $y(t_0 + nh)$ for $n \geq k$ can then be computed recursively from (1.1). In the case $\beta_k = 0$ we have an explicit method, otherwise it is implicit and the numerical solution y_{n+k} has to be computed iteratively.

Germund Dahlquist¹

Since the fundamental work of Dahlquist (1956) it is common to denote the generating polynomials of the coefficients by

$$\rho(\zeta) = \sum_{j=0}^k \alpha_j \zeta^j, \quad \sigma(\zeta) = \sum_{j=0}^k \beta_j \zeta^j.$$

For the classical theory of multistep methods we refer the reader to Chap. III of Hairer, Nørsett & Wanner (1993). We just recall some important definitions.

Order. A multistep method has order r if, when applied with exact starting values to the problem $\dot{y} = t^q$ ($0 \leq q \leq r$), it integrates the problem without error. This is equivalent to the requirement that

$$\rho(e^h) - h\sigma(e^h) = \mathcal{O}(h^{r+1}) \quad \text{for } h \rightarrow 0. \quad (1.2)$$

Stability. Method (1.1) is stable if, when applied to $\dot{y} = 0$, it yields for all y_0, \dots, y_{k-1} a bounded numerical solution. This is equivalent to the requirement that the polynomial $\rho(\zeta)$ satisfies the root condition, i.e., all roots of $\rho(\zeta) = 0$ satisfy $|\zeta| \leq 1$, and those on the unit circle are simple roots. The method is called *strictly stable*, if all roots are inside the unit circle with the exception of $\zeta = 1$.

Convergence. If a multistep method is stable and of order $r \geq 1$, it is convergent of order r for all sufficiently smooth problems. This means that, assuming starting approximations with an error bounded by $\mathcal{O}(h^r)$, the global error satisfies $y_n - y(t_0 + nh) = \mathcal{O}(h^r)$ on compact intervals $nh \leq T$.

Symmetry. If the coefficients of a multistep formula (1.1) satisfy

$$\alpha_{k-j} = -\alpha_j, \quad \beta_{k-j} = \beta_j \quad \text{for all } j, \quad (1.3)$$

then the method is called symmetric. Condition (1.3) implies that for every zero ζ of $\rho(\zeta)$ also its inverse ζ^{-1} is a zero. Hence, for stable symmetric methods all zeros of $\rho(\zeta)$ are simple and lie on the unit circle.

Example 1.1. We consider the pendulum equation (I.1.13), and we apply the following multistep methods: the 2-step explicit Adams method

$$y_{n+2} = y_{n+1} + h \left(\frac{3}{2} f_{n+1} - \frac{1}{2} f_n \right), \quad (1.4)$$

the 2-step backward differentiation formula (BDF)

$$\frac{3}{2} y_{n+2} - 2y_{n+1} + \frac{1}{2} y_n = h f_{n+2}, \quad (1.5)$$

and the (2-step) symmetric explicit midpoint rule

¹ Germund Dahlquist, born: 16 January 1925 in Uppsala (Sweden), died: 8 February 2005.

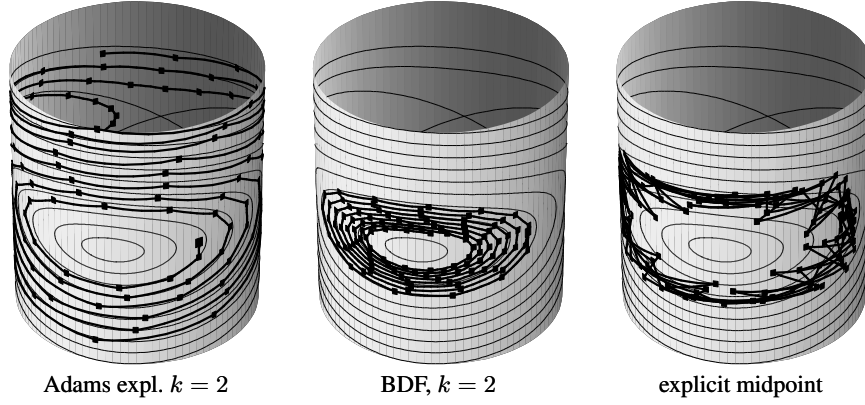


Fig. 1.1. Solutions of the pendulum problem (I.1.13); explicit Adams with step size $h = 0.5$, initial value $(p_0, q_0) = (0, 0.7)$; BDF with step size $h = 0.5$, initial value $(p_0, q_0) = (0, 0.95)$; explicit midpoint rule with $h = 0.4$ and initial value $(p_0, q_0) = (1.1, 0)$

$$y_{n+2} = y_n + 2hf_{n+1}. \quad (1.6)$$

For all methods we take $y_1 = y_0 + hf_0$ as the approximation for $y(t_0 + h)$. The results of the first 108 steps are shown in Fig. 1.1. We observe that the first two methods, as expected, behave similarly as the explicit and implicit Euler method (the numerical solution spirals either outwards or inwards). This will be rigorously explained in Sect. XV.2.1 below. However, as might not be expected, the symmetric method (1.6) does not behave like the implicit midpoint rule (cf. Fig. I.1.4), it shows undesired increasing oscillations (parasitic solutions).

After this negative experience with classical multistep methods, the obvious question is: are there multistep methods which have a long-time behaviour that is comparable to symplectic and/or symmetric one-step methods?

XV.1.2 Multistep Methods for Second Order Equations

Many important Hamiltonian systems are second order differential equations

$$\ddot{y} = f(y), \quad (1.7)$$

where the force f is independent of the velocity \dot{y} . Introducing the new variable $v = \dot{y}$, we obtain the system $\dot{y} = v$, $\dot{v} = f(y)$ of first order equations. If we apply a multistep method (1.1) with generating polynomials $\rho^*(\zeta) = \sum_{j=0}^{k^*} \alpha_j^* \zeta^j$ and $\sigma^*(\zeta) = \sum_{j=0}^{k^*} \beta_j^* \zeta^j$ to this system, we get

$$\sum_{j=0}^{k^*} \alpha_j^* y_{n+j} = h \sum_{j=0}^{k^*} \beta_j^* v_{n+j}, \quad \sum_{j=0}^{k^*} \alpha_j^* v_{n+j} = h \sum_{j=0}^{k^*} \beta_j^* f(y_{n+j}).$$

An elimination of the v -variables then yields

$$\sum_{j=0}^k \alpha_j y_{n+j} = h^2 \sum_{j=0}^k \beta_j f(y_{n+j}), \quad (1.8)$$

where $k = 2k^*$, $\rho(\zeta) = \rho^*(\zeta)^2$ and $\sigma(\zeta) = \sigma^*(\zeta)^2$. We consider here methods (1.8) which do not necessarily originate from a multistep method for first order equations, and we denote the generating polynomials of the coefficients α_j and β_j again by $\rho(\zeta)$ and $\sigma(\zeta)$. From the classical theory (see Sect. III.10 of Hairer, Nørsett & Wanner 1993) we recall the following definitions and results.

Order. A method (1.8) has order r if its generating polynomials satisfy

$$\rho(e^h) - h^2 \sigma(e^h) = \mathcal{O}(h^{r+2}) \quad \text{for } h \rightarrow 0. \quad (1.9)$$

Stability. Method (1.8) is stable if all zeros of the polynomial $\rho(\zeta)$ satisfy $|\zeta| \leq 1$, and those on the unit circle are at most double zeros. Observe that for methods originating from (1.1) all zeros are double. The method is called *strictly stable*, if all zeros are inside the unit circle with the exception of $\zeta = 1$.

Convergence. If a multistep method (1.8) is stable, of order $r \geq 1$ and if the starting values are accurate enough, the global error satisfies $y_n - y(t_0 + nh) = \mathcal{O}(h^r)$ on compact intervals $nh \leq T$.

Symmetry. If the coefficients of (1.8) satisfy

$$\alpha_{k-j} = \alpha_j, \quad \beta_{k-j} = \beta_j \quad \text{for all } j, \quad (1.10)$$

then the method is symmetric. Again, for every zero ζ of $\rho(\zeta)$ the value ζ^{-1} is also a zero. Hence, stable symmetric methods have all zeros of $\rho(\zeta)$ on the unit circle and they are at most of multiplicity two.

Dahlquist (1956) noticed that double zeros of $\rho(\zeta)$ on the unit circle can lead to an exponential error growth. Lambert & Watson (1976) analyzed in detail the application of (1.8) to the linear test equation $\ddot{y} = -\omega^2 y$. They found that with symmetric methods for which $\rho(\zeta)$ does not have double roots on the unit circle other than $\zeta = 1$, the numerical solution remains close to a periodic orbit (for sufficiently small step sizes). For example, the Störmer–Verlet method $y_{n+1} - 2y_n + y_{n-1} = h^2 f_n$ satisfies this property for $0 < h\omega < 2$ (see Sect. I.5.2). The study of the long-time behaviour of symmetric methods (1.8) was then put forward by the article of Quinlan & Tremaine (1990), where an excellent performance for simulations of the outer solar system is reported.

Example 1.2. We consider the Kepler problem (I.2.2) with initial values (I.2.11) and eccentricity $e = 0.2$. We apply the following three methods with constant step size $h = 0.01$ on the interval of length $2\pi \cdot 10^5$ (i.e., 10^5 periods):

- (A) $y_{n+4} - 2y_{n+3} + y_{n+2} = h^2 \left(\frac{7}{6} f_{n+3} - \frac{5}{12} f_{n+2} + \frac{1}{3} f_{n+1} - \frac{1}{12} f_n \right)$
- (B) $y_{n+4} - 2y_{n+2} + y_n = h^2 \left(\frac{4}{3} f_{n+3} + \frac{4}{3} f_{n+2} + \frac{4}{3} f_{n+1} \right)$
- (C) $y_{n+4} - 2y_{n+3} + 2y_{n+2} - 2y_{n+1} + y_n = h^2 \left(\frac{7}{6} f_{n+3} - \frac{1}{3} f_{n+2} + \frac{7}{6} f_{n+1} \right).$

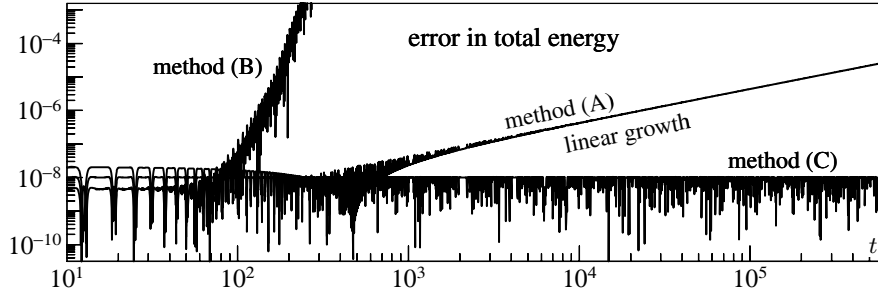


Fig. 1.2. Error in the total energy for the three linear multistep methods of Example 1.2 applied to the Kepler problem with $e = 0.2$

All three methods are of order $r = 4$; method (A) is strictly stable, whereas methods (B) and (C) are symmetric. For method (B) the ρ -polynomial has a double root at $\zeta = -1$, for method (C) it does not have double roots other than 1. Starting values y_1, y_2 , and y_3 are computed very accurately with a high-order Runge–Kutta method.

The error in the total energy is plotted for all three methods in Fig. 1.2. On the first 10 periods, all methods behave similarly and no error growth is observed. Beyond this interval, method (A) shows a linear error growth (as it is the case for non-symplectic and non-symmetric one-step methods), method (B) has an exponential error growth, and for method (C) the error remains bounded of size $\mathcal{O}(h^4)$ on the whole interval of integration. One of the aims of this chapter is to explain the excellent long-time behaviour of method (C).

Stabilized Version of (1.8). Due to the double zeros (of modulus one) of the characteristic polynomial of the difference equation $\sum_j \alpha_j y_{n+j} = 0$, we have an undesired propagation of rounding errors (especially for long-time integrations). To overcome this difficulty, we split the characteristic polynomial $\rho(\zeta)$ into

$$\rho(\zeta) = \rho_A(\zeta) \cdot \rho_B(\zeta), \quad (1.11)$$

such that each polynomial

$$\rho_A(\zeta) = \sum_{j=0}^{k_A} \alpha_j^{(A)} \zeta^j, \quad \rho_B(\zeta) = \sum_{j=0}^{k_B} \alpha_j^{(B)} \zeta^j$$

has only simple roots of modulus one. Introducing the new variable $h v_n := \sum_j \alpha_j^{(A)} y_{n+j}$, the recurrence relation (1.8) becomes equivalent to

$$\sum_{j=0}^{k_A} \alpha_j^{(A)} y_{n+j} = h v_n, \quad \sum_{j=0}^{k_B} \alpha_j^{(B)} v_{n+j} = h \sum_{j=0}^k \beta_j f_{n+j}. \quad (1.12)$$

This formula, which for the Störmer–Verlet scheme corresponds to the one-step formulation (I.1.17), is much better suited for an implementation. If the splitting is such that $\rho'_A(1) = 1$, the discretization (1.12) is consistent with the first order partitioned system $\dot{y} = v, \dot{v} = f(y)$.

XV.1.3 Partitioned Multistep Methods

Motivated by the stabilized version (1.12) of multistep methods for second order equations, let us consider general partitioned systems of differential equations

$$\dot{y} = f(y, v), \quad \dot{v} = g(y, v), \quad (1.13)$$

where, needless to say, y and v may be vectors. The idea is to apply different multistep methods to different components. We thus get

$$\sum_{j=0}^k \alpha_j^{(A)} y_{n+j} = h \sum_{j=0}^k \beta_j^{(A)} f_{n+j}, \quad \sum_{j=0}^k \alpha_j^{(B)} v_{n+j} = h \sum_{j=0}^k \beta_j^{(B)} g_{n+j}, \quad (1.14)$$

where $f_n = f(y_n, v_n)$ and $g_n = g(y_n, v_n)$. We can take the same k for both methods without loss of generality, if we abandon the assumption $|\alpha_0| + |\beta_0| > 0$.

Such a method is of order r , if both methods are of order r . It is stable (strictly stable, symmetric, . . .), if both methods are stable (strictly stable, symmetric, . . .).

Example 1.3. For our next experiment we use the symmetric methods

$$\begin{aligned} \text{(A)} : \quad & y_{n+3} - y_{n+2} + y_{n+1} - y_n = h(f_{n+2} + f_{n+1}) \\ \text{(B)} : \quad & v_{n+3} - v_{n+1} = 2hg_{n+2}. \end{aligned} \quad (1.15)$$

Both methods are of order 2, and their ρ -polynomials $\rho_A(\zeta) = (\zeta - 1)(\zeta^2 + 1)$ and $\rho_B(\zeta) = (\zeta - 1)(\zeta + 1)$ do not have common zeros with the exception of $\zeta = 1$.

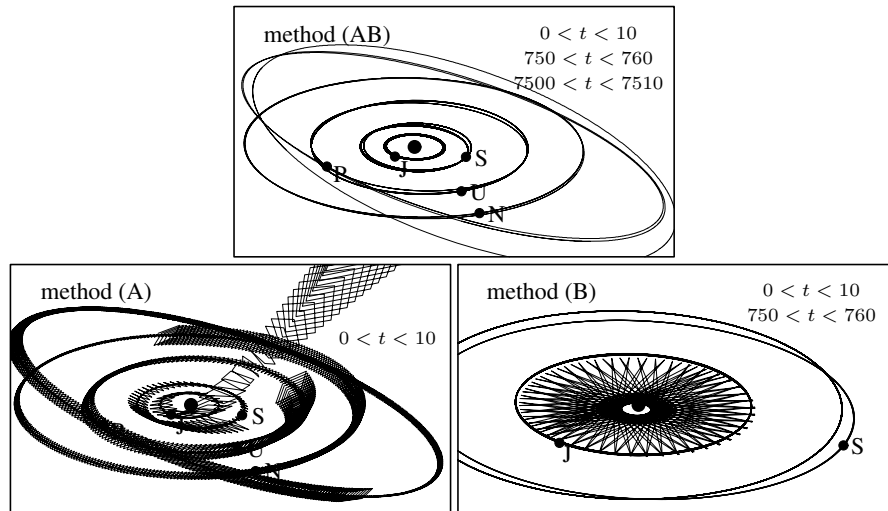
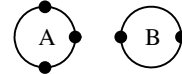


Fig. 1.3. Three versions of the methods (1.15) applied with step size $h = 50$ (days) to the outer solar system. For method (B) only the numerical orbits of Jupiter and Saturn are plotted. The time intervals are given in units of 10 000 days

We choose the outer solar system with the data as described in Sect. I.2.4, and we apply the methods in three versions: (i) as partitioned method (AB), where the positions are treated by method (A) and the velocities by method (B); (ii) method (A) is applied to all components; (iii) method (B) is applied to all components. The numerical results are shown in Fig. 1.3. Whereas the individual methods show instabilities on rather short time intervals, the partitioned method gives a correct picture even with a large step size $h = 50$.

XV.2 The Underlying One-Step Method

Much insight into the long-time behaviour of multistep methods can be gained by relating their numerical solution to one-step methods. This then allows for an application of the considerations of the preceding sections.

XV.2.1 Strictly Stable Multistep methods

It was a surprising result when Kirchgraber (1986) proved that strictly stable multistep methods are essentially equivalent to one-step methods. Although this one-step method is “quite exotic” (Eirola & Nevanlinna 1988), it is the key for a better understanding of the dynamics of strictly stable methods.

Theorem 2.1 (Kirchgraber 1986). *Consider a strictly stable linear multistep method (1.1) applied with a sufficiently small step size h . Then, there exists a one-step method Φ_h such that for starting approximations computed by $y_j = \Phi_h^j(y_0)$, $j = 1, \dots, k-1$, the numerical solution of (1.1) is identical to that obtained by the one-step method, i.e., $y_{n+1} = \Phi_h(y_n)$ for all $n \geq 0$.*

Proof. The idea is to reformulate the multistep method (1.1) in such a way that the Invariant Manifold Theorem of Sect. XII.3 can be applied. To keep the notation as simple as possible, let us consider the case $k = 3$.

We write the method in the form

$$\begin{pmatrix} y_{n+3} \\ y_{n+2} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} -a_2 & -a_1 & -a_0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y_{n+2} \\ y_{n+1} \\ y_n \end{pmatrix} + h \begin{pmatrix} F_h(y_n, y_{n+1}, y_{n+2}) \\ 0 \\ 0 \end{pmatrix} \quad (2.1)$$

with $a_i = \alpha_i/\alpha_k$, and we transform the appearing matrix A to Jordan canonical form $J = T^{-1}AT$. We thus get

$$Z_{n+1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & d_{11} & d_{12} \\ 0 & d_{21} & d_{22} \end{pmatrix} Z_n + hG(Z_n), \quad Z_n = T^{-1} \begin{pmatrix} y_{n+2} \\ y_{n+1} \\ y_n \end{pmatrix}. \quad (2.2)$$

Since the method is strictly stable, 1 is a simple eigenvalue of A , and all other eigenvalues are less than 1 in modulus. Consequently, the matrix $D = (d_{ij})$ satisfies

$\|D\| < 1$ in a suitable norm. Partitioning $Z_n = (\xi_n, \eta_n)^T$ into its first component ξ_n and the rest (collected in η_n), we see that (2.2) is of the form (XII.3.1) with L_{xx}, L_{xy}, L_{yx} of size $\mathcal{O}(h)$, and $L_{yy} = \|D\| < 1$. Theorem XII.3.1 thus yields the existence of a function $\eta = s(\xi)$ such that the manifolds

$$\mathcal{N}_h = \left\{ \begin{pmatrix} \xi \\ s(\xi) \end{pmatrix} ; \xi \in \mathbb{R}^d \right\} \quad \text{and} \quad \mathcal{M}_h = \left\{ T \begin{pmatrix} \xi \\ s(\xi) \end{pmatrix} ; \xi \in \mathbb{R}^d \right\}$$

are invariant under the mappings (2.2) and (2.1), respectively. The function $s(\xi)$ is Lipschitz continuous with constant $\lambda = \mathcal{O}(h)$.

Since the first column of T , which is the eigenvector corresponding to the eigenvalue 1 of A , is given by $(1, 1, 1)^T$, the last component of $T \begin{pmatrix} \xi \\ s(\xi) \end{pmatrix}$ satisfies $y = \xi + g(\xi)$ where $g(\xi)$ is Lipschitz continuous with constant $\mathcal{O}(h)$. By the Banach fixed-point theorem this equation has a unique solution $\xi = r(y)$. Consequently, the manifold \mathcal{M}_h can be parametrized in terms of y as

$$\mathcal{M}_h = \{(\Psi_h(y), \Phi_h(y), y)^T ; y \in \mathbb{R}^d\}.$$

Its invariance under (2.1) implies that

$$\begin{pmatrix} \Psi_h(\hat{y}) \\ \Phi_h(\hat{y}) \\ \hat{y} \end{pmatrix} = \begin{pmatrix} -a_2 & -a_1 & -a_0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \Psi_h(y) \\ \Phi_h(y) \\ y \end{pmatrix} + h \begin{pmatrix} F_h(y, \Phi_h(y), \Psi_h(y)) \\ 0 \\ 0 \end{pmatrix}$$

and consequently $\hat{y} = \Phi_h(y)$ and $\Phi_h(\hat{y}) = \Psi_h(y)$, so that $\Psi_h(y) = \Phi_h^2(y)$. This holds for all y , and thus proves the statement of the theorem. \square

Example 2.2. For a scalar linear problem $\dot{y} = \lambda y$, the application of a multistep method yields a difference equation with characteristic polynomial $\rho(\zeta) - h\lambda\sigma(\zeta)$. Denoting its zeros by $\zeta_1(h\lambda), \dots, \zeta_k(h\lambda)$, where $\zeta_1(0) = 1$ and $|\zeta_j(0)| < 1$ for $j \geq 2$, the numerical solution can be written as (assuming distinct $\zeta_j(h\lambda)$)

$$y_n = c_1 \zeta_1^n(h\lambda) + c_2 \zeta_2^n(h\lambda) + \dots + c_k \zeta_k^n(h\lambda).$$

The coefficients c_1, \dots, c_k depend on $h\lambda$ and are determined by the starting approximations y_0, \dots, y_{k-1} . In this situation the underlying one-step method is the mapping $y_0 \mapsto \zeta_1(h\lambda)y_0$. Observe that $\zeta_1(z)$ is in general not a rational function as we are used to with Runge–Kutta methods.

Remark 2.3 (Asymptotic Phase). For arbitrary y_0, y_1, \dots, y_{k-1} close to the exact solution, there exists y_0^* such that the multistep solution $\{y_n\}$ and the one-step solution $\{y_n^*\}$, given by $y_{n+1}^* = \Phi_h(y_n^*)$, approach exponentially fast, i.e.,

$$\|y_n - y_n^*\| \leq \text{Const} \cdot \rho^n \quad \text{for all } n \geq 0 \quad (2.3)$$

with some ρ satisfying $0 < \rho < 1$ (see Exercise XII.3). This is due to the attractivity of the invariant manifold \mathcal{M}_h . A proof is given in Stoffer (1993), and it is based on techniques of Nipp & Stoffer (1992). This result explains why strictly stable linear multistep methods have the same long-time behaviour as one-step methods.

In the context of “geometric numerical integration” we are mainly interested in symplectic and/or symmetric methods which, for linear problems, are characterized by the condition $\zeta_1(-z)\zeta_1(z) \equiv 1$ (see Sect. VI.4.2). This, however, is only possible for symmetric multistep methods (Exercise 1) which cannot be strictly stable.

XV.2.2 Formal Analysis for Weakly Stable Methods

The proof and the statement of Theorem 2.1 break down as soon as at least one root of $\rho(\zeta)$, different from 1, has modulus one. Moreover, Example 2.2 shows that we cannot expect a property like (2.3) with $\rho < 1$. All we can hope for is to find an underlying one-step method as a formal series in h . Surprisingly, this provides a lot of insight into the long-time dynamics of weakly stable multistep methods.

Theorem 2.4. *Consider a linear multistep method (1.1), and assume that $\zeta = 1$ is a single root of $\rho(\zeta) = 0$. Then there exists a unique formal expansion*

$$\Phi_h(y) = y + h d_1(y) + h^2 d_2(y) + \dots \quad (2.4)$$

such that

$$\sum_{j=0}^k \alpha_j \Phi_h^j(y) = h \sum_{j=0}^k \beta_j f(\Phi_h^j(y)),$$

where identity is understood in the sense of formal power series in h .

If the multistep method is of order r , then also the underlying one-step method is of order r , i.e., $\Phi_h(y) - \varphi_h(y) = \mathcal{O}(h^{r+1})$.

The formal series for $\Phi_h(y)$ is called “step-transition operator” in the Chinese literature (see e.g., Feng (1995), page 274). We call it “underlying one-step method”. Notice that this theorem does not require any stability assumption.

Proof. Expanding $\Phi_h^j(y)$ and $f(\Phi_h^j(y))$ into powers of h , a comparison of the coefficients yields

$$\begin{aligned} \rho'(1) d_1(y) &= \sigma(1) f(y) \\ \rho'(1) d_2(y) &= -\frac{\rho''(1)}{2} d_1'(y) d_1(y) + \sigma'(1) f'(y) d_1(y) \\ \rho'(1) d_j(y) &= \dots, \end{aligned} \quad (2.5)$$

where the three dots represent known functions depending on derivatives of $f(y)$ and on $d_i(y)$ with $i < j$. Since $\rho'(1) \neq 0$ by assumption, unique coefficient functions $d_j(y)$ are obtained recursively. The statement on the order follows from the fact that the exact flow $\varphi_h(y)$ has a defect $\mathcal{O}(h^{r+1})$ in the multistep formula. \square

The computation of the previous proof shows that the series (2.4) is a B-series. This follows rigorously from the results of Sect. III.1.4. Whereas the B-series representation of Runge–Kutta methods converges for sufficiently small h , this is in general not the case for (2.4); see the next example.

Example 2.5. Consider a consistent two-step method

$$\alpha_2 y_{n+2} + \alpha_1 y_{n+1} + \alpha_0 y_n = h(\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n),$$

and apply it to the simple system $\dot{y} = f(t)$, $\dot{t} = 1$. The y -component of the underlying one-step method then takes the form

$$\Phi_h(t_0, y_0) = y_0 + \sum_{j \geq 1} h^j a_j f^{(j-1)}(t_0). \quad (2.6)$$

Putting $f(t) = e^t$ yields

$$A(\zeta) = \sum_{j \geq 1} a_j \zeta^{j-1} = \frac{\beta_2 e^{2\zeta} + \beta_1 e^\zeta + \beta_0}{\alpha_2(1 + e^\zeta) + \alpha_1}.$$

for the generating function of the coefficients a_j . Since this function has finite poles, the radius of convergence of $A(\zeta)$ is finite. Therefore, the radius of convergence of the series (2.6) has to be zero as soon as $f^{(j)}(t_0)$ behaves like $j! \mu \kappa^j$ (this is typically the case for analytic functions). Independent of the fact whether the method is strictly stable or not, the series (2.6) usually does not converge.

Both, Theorem 2.1 and Theorem 2.4, extend in a straightforward manner to partitioned multistep methods (1.14). To get analogous results for multistep methods (1.8) for second order differential equations, one has to introduce an approximation for the velocity $v = \dot{y}$. This will be explained in more detail in Sect. XV.3 below.

XV.3 Backward Error Analysis

The backward error analysis for multistep methods (Hairer 1999) is presented in two steps:

- for “smooth” numerical solutions (obtained by the underlying one-step method);
- for the general case.

The idealized situation of no parasitic terms gives already much insight into conservation properties of the method (see Sect. XV.4). The study of the general case is, however, necessary for getting estimates for the parasitic solutions (Sect. XV.5), so that rigorous statements on the long-time behaviour are possible.

XV.3.1 Modified Equation for Smooth Numerical Solutions

The formal backward error analysis of Chap. IX could be directly applied to the underlying one-step method of Sect. XV.2.2. However, due to the non-convergence of the series for $\Phi_h(y)$, difficulties may arise as soon as rigorous estimates are desired. We prefer to derive the modified differential equation directly from the multistep formula and thus avoid the use of the underlying one-step method.

Theorem 3.1. Consider a linear multistep method (1.1), and assume that $\rho(1) = 0$ and $\rho'(1) = \sigma(1) \neq 0$. Then there exist unique h -independent functions $f_j(y)$ such that, for every truncation index N , every solution of

$$\dot{y} = f(y) + hf_2(y) + h^2f_3(y) + \dots + h^{N-1}f_N(y) \quad (3.1)$$

satisfies

$$\sum_{j=0}^k \alpha_j y(t+jh) = h \sum_{j=0}^k \beta_j f(y(t+jh)) + \mathcal{O}(h^{N+1}). \quad (3.2)$$

If the multistep method is of order r , then $f_j(y) = 0$ for $2 \leq j \leq r$. If the method is symmetric, then $f_j(y) = 0$ for all even j , so that the modified equation (3.1) has an expansion in even powers of h .

Proof. Using the Lie derivative $(D_i g)(y) = g'(y)f_i(y)$ (with $f_1(y) = f(y)$) and $D = D_1 + hD_2 + h^2D_3 + \dots$, the solution of (3.1) with initial value $y(t) = y$ satisfies $y(t+jh) = e^{jhD}y + \mathcal{O}(h^{N+1})$ and $f(y(t+jh)) = e^{jhD}f(y) + \mathcal{O}(h^{N+1})$ (by Taylor expansion). We thus have

$$\rho(e^{hD})y = h\sigma(e^{hD})f(y) + \mathcal{O}(h^{N+1}). \quad (3.3)$$

With the expansion $x\sigma(e^x)/\rho(e^x) = 1 + \mu_1x + \mu_2x^2 + \dots$ this becomes

$$\dot{y} = (1 + \mu_1hD + \mu_2h^2D^2 + \dots)f(y) + \mathcal{O}(h^N). \quad (3.4)$$

A comparison with (3.1) yields $f_1(y) = f(y)$, and

$$f_j(y) = \sum_{l \geq 1} \mu_l \sum_{j_1 + \dots + j_l = j-1} (D_{j_1} \dots D_{j_l} f)(y) \quad (3.5)$$

for $j \geq 2$, which uniquely defines the functions $f_j(y)$ in a recursive manner. \square

Lemma 3.2. If $f(y)$ is analytic and bounded by M in $B_R(y_0)$, then we have

$$\|f_j(y)\| \leq \mu M \left(\frac{\eta M j}{R} \right)^{j-1} \quad \text{for} \quad \|y - y_0\| \leq R/2, \quad (3.6)$$

where μ and η depend only on the coefficients α_j, β_j of the multistep method.

Proof. The estimate (3.6) is obtained as in the proof of Theorem IX.7.5. We just sketch the main idea in the notation used there. With $\delta = R/(2(J-1))$ we have $\|f_j\|_j \leq \delta b_j$, where the generating function $b(\zeta) = \sum_{j \geq 1} b_j \zeta^j$ of the b_j satisfies

$$b(\zeta) = \frac{M\zeta}{\delta} \left(1 + \sum_{l \geq 1} |\mu_l| b(\zeta)^l \right).$$

By the implicit function theorem, $b(\zeta)$ is analytic and bounded in a disc of radius $c\delta/M$ centred at the origin (c is a positive constant depending only on the coefficients of the multistep method). The estimate (3.6) then follows from Cauchy's inequalities as in the proof of Theorem IX.7.5. \square

It is remarkable that, although the Taylor series of the underlying one-step method generally diverges, the coefficient functions of the modified differential equation satisfy the same estimate as for Runge–Kutta methods. This enables us to prove an analogue of Theorem IX.7.6 which, for one-step methods, is the main ingredient for exponentially small error estimates. One can prove that for suitably chosen $N = N(h)$ and for $h \leq h_0/4$ with $h_0 = R/(e\eta M)$, the solution of (3.1) satisfies

$$\left\| \sum_{j=0}^k \alpha_j y(t+jh) - h \sum_{j=0}^k \beta_j f(y(t+jh)) \right\| \leq h\gamma M e^{-h_0/h},$$

where γ depends only on the multistep formula. The proof of this statement is similar to that of Theorem IX.7.6. We skip details and refer to Hairer (1999).

For strictly stable multistep methods, Theorem 2.1 together with the Invariant Manifold Theorem XII.3.2 thus imply that the underlying one-step method is exponentially close to the exact solution of the truncated modified equation. The parasitic solution terms are rapidly damped out by the property (2.3) of asymptotic phase. The same conclusions as for one-step methods can therefore be drawn.

For symmetric methods the situation is not so simple. One has to study the parasitic solution components to get information on the long-time behaviour of the numerical solution of (1.1). The basic techniques will be prepared in Sect. XV.3.2.

Partitioned Multistep Methods. The extension of the modified differential equation to methods (1.14) is straightforward. There exist functions $f_j(y, v)$ and $g_j(y, v)$ such that the exact solution of

$$\begin{aligned} \dot{y} &= f(y, v) + hf_2(y, v) + \dots + h^{N-1}f_N(y, v) \\ \dot{v} &= g(y, v) + hg_2(y, v) + \dots + h^{N-1}g_N(y, v) \end{aligned} \quad (3.7)$$

satisfies the multistep formula (1.14) up to a defect of size $\mathcal{O}(h^{N+1})$. The coefficient functions can be computed by comparing (3.7) to

$$\begin{aligned} \dot{y} &= (1 + \mu_1^{(A)}hD + \mu_2^{(A)}h^2D^2 + \dots)f(y, v) + \mathcal{O}(h^N) \\ \dot{v} &= (1 + \mu_1^{(B)}hD + \mu_2^{(B)}h^2D^2 + \dots)g(y, v) + \mathcal{O}(h^N), \end{aligned} \quad (3.8)$$

where the real numbers $\mu_j^{(A)}$ and $\mu_j^{(B)}$ are given by $x\sigma^{(A)}(e^x)/\rho^{(A)}(e^x) = 1 + \mu_1^{(A)}x + \mu_2^{(A)}x^2 + \dots$ and by $x\sigma^{(B)}(e^x)/\rho^{(B)}(e^x) = 1 + \mu_1^{(B)}x + \mu_2^{(B)}x^2 + \dots$, respectively. The Lie operator is defined by $D = D_1 + hD_2 + h^2D_3 + \dots$, where $(D_j\Psi)(y, v) = \Psi_y(y, v)f_j(y, v) + \Psi_v(y, v)g_j(y, v)$, and it corresponds to the time derivative of solutions of (3.7).

Multistep Methods for Second Order Differential Equations. The method (1.8) for differential equations $\ddot{y} = f(y)$ can be treated in a similar way. In the absence of derivative approximations we get a modified differential equation of the second order

$$\ddot{y} = f(y) + hf_2(y, \dot{y}) + \dots + h^{N-1}f_N(y, \dot{y}), \quad (3.9)$$

where the perturbation terms also depend on \dot{y} . Its exact solution satisfies the multistep relation (1.8) up to a defect of size $\mathcal{O}(h^{N+2})$, if (3.9) is equivalent to

$$\ddot{y} = (1 + \mu_1 hD + \mu_2 h^2 D^2 + \dots)f(y) + \mathcal{O}(h^N), \quad (3.10)$$

where $x^2\sigma(e^x)/\rho(e^x) = 1 + \mu_1 x + \mu_2 x^2 + \dots$, and the time derivative is given by the Lie operator $D = D_1 + hD_2 + h^2 D_3 + \dots$ with $(D_1\Psi)(y, \dot{y}) = \Psi_y(y, \dot{y})\dot{y} + \Psi_{\dot{y}}(y, \dot{y})f(y)$ and $(D_j\Psi)(y, \dot{y}) = \Psi_{\dot{y}}(y, \dot{y})f_j(y, \dot{y})$ for $j \geq 2$. A comparison of equal powers of h in (3.9) and (3.10) uniquely defines the coefficient functions $f_j(y, \dot{y})$.

If the multistep method (1.8) is complemented with a difference formula for approximations of the derivative $v = \dot{y}$ at grid points,

$$v_n = \frac{1}{h} \sum_{j=-l}^l \delta_j y_{n+j}, \quad (3.11)$$

we get an additional modified differential equation

$$v = (1 + \nu_1 hD + \nu_2 h^2 D^2 + \dots)\dot{y}. \quad (3.12)$$

The coefficients ν_j are given by $x^{-1}\delta(e^x) = 1 + \nu_1 x + \nu_2 x^2 + \dots$, where $\delta(\zeta) = \sum_{j=-l}^l \delta_j \zeta^j$. For given y , this relation gives a formal one-to-one correspondence between v and \dot{y} . Consequently, the differential equation (3.10) combined with (3.12) can be considered as a first order differential system for the variables y and v .

XV.3.2 Parasitic Modified Equations

In practice, due to the necessity of starting approximations y_1, \dots, y_{k-1} , the numerical solution of a multistep method does not lie on a solution of (3.1). For methods, where initial perturbations are not damped out sufficiently fast (cf. property (2.3) of asymptotic phase), an additional investigation is therefore needed for the study of the propagation of perturbations in the starting approximations. Let us start with two illustrating numerical experiments.

Example 3.3. Consider the explicit, linear 3-step method

$$y_{n+3} - y_{n+2} + y_{n+1} - y_n = h(f_{n+2} + f_{n+1}), \quad (3.13)$$

with characteristic polynomial $\rho(\zeta) = (\zeta - 1)(\zeta^2 + 1)$, and apply it to the pendulum equation (I.1.13). For a better illustration of the propagation of errors we consider starting approximations y_1, y_2 that are rather far from the exact solution passing through y_0 . The result is shown in Fig. 3.1. We observe that the numerical solution does not lie on one smooth curve, but on four curves, and every fourth solution approximation is on the same curve.

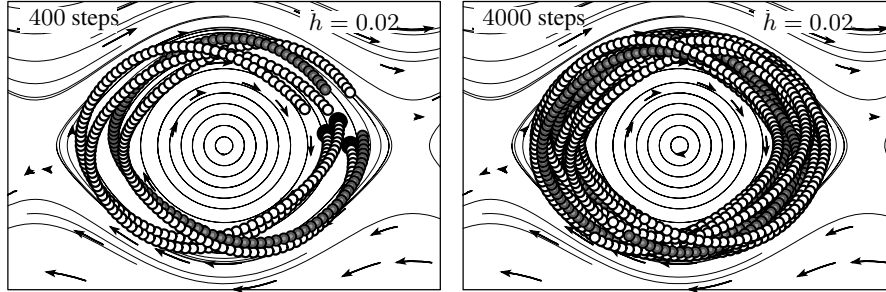


Fig. 3.1. Numerical solution of (3.13) applied to the pendulum equation. The initial approximations $y_0 = (1.9, 0.4)$, $y_1 = (1.7, 0.2)$, $y_2 = (2.1, 0)$ are indicated by black bullets; the solution points y_3, y_7, y_{11}, \dots in grey

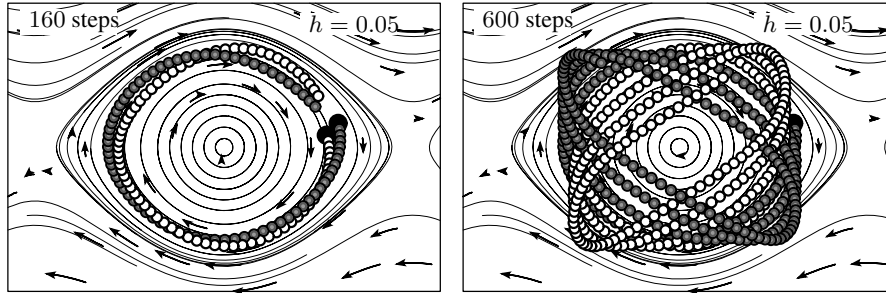


Fig. 3.2. Numerical solution of the explicit midpoint rule (3.14) applied to the pendulum equation. The initial approximations $y_0 = (1.9, 0.4)$, $y_1 = (1.7, 0.2)$ are indicated by black bullets; the solution points y_2, y_4, y_6, \dots in grey

This example shows an unexpected good long-time behaviour. Although the starting approximations are far from a smooth solution, the distance of the numerical approximations to a smooth solution curve does not increase. This is, however, not the typical situation as can be seen from our next experiment.

Example 3.4. We consider the explicit midpoint rule

$$y_{n+2} - y_n = 2h f_{n+1}, \quad (3.14)$$

which has $\rho(\zeta) = (\zeta - 1)(\zeta + 1)$ as characteristic polynomial. This time, the numerical solution (see Fig. 3.2) lies on two smooth curves. In contrast to the previous example, an unacceptable linear growth of the perturbations can be observed.

To be able to explain this behaviour of the multistep solutions, we complement the analysis of the modified equation for smooth numerical solutions with so-called parasitic modified equations. This theory has been developed by Hairer (1999) for first order differential equations, and extended to second order systems by Hairer & Lubich (2004).

Consider a stable, symmetric multistep method (1.1) and denote the zeros of its characteristic polynomial $\rho(\zeta)$ by $\zeta_1 = 1$ (principal root) and ζ_2, \dots, ζ_k (parasitic roots). We then enumerate the set of all finite products,

$$\{\zeta_\ell\}_{\ell \in \mathcal{I}} = \{\zeta = \zeta_1^{m_1} \cdots \zeta_k^{m_k} ; m_j \geq 0\} = \{\zeta_1, \dots, \zeta_k, \zeta_{k+1}, \dots\}. \quad (3.15)$$

It is $\{1, i, -i, -1\}$ for method (3.13) and $\{1, -1\}$ for the explicit midpoint rule (3.14). The set of subscripts \mathcal{I} can be finite or infinite. We let $\mathcal{I}^* = \mathcal{I} \setminus \{1\}$, and we denote by \mathcal{I}_N^* and \mathcal{I}_N the finite subsets of elements which, in the representation (3.15), have $\sum_j m_j < N$.

Motivated by the previous examples and by representations of the asymptotic expansion of the global error of weakly stable multistep methods (see for example Sect. III.9 of Hairer, Nørsett & Wanner, 1993), we aim at writing the general solution y_n of the multistep method (1.1) in the form

$$y_n = y(nh) + \sum_{\ell \in \mathcal{I}^*} \zeta_\ell^n z_\ell(nh), \quad (3.16)$$

where $y(t)$ and $z_\ell(t)$ are smooth functions (with derivatives bounded independently of h). The following result extends Theorem 3.1.

Theorem 3.5. *Consider a stable, consistent, and symmetric multistep method (1.1). For every truncation index $N \geq 2$, there then exist h -independent functions $f_{\ell,j}(y, \mathbf{z}^*)$ with $\mathbf{z}^* = (z_\ell)_{\ell=2}^k$ such that for every solution of*

$$\begin{aligned} \dot{y} &= f_{1,1}(y, \mathbf{z}^*) + hf_{1,2}(y, \mathbf{z}^*) + \dots + h^{N-1}f_{1,N}(y, \mathbf{z}^*) \\ \dot{z}_\ell &= f_{\ell,1}(y, \mathbf{z}^*) + hf_{\ell,2}(y, \mathbf{z}^*) + \dots + h^{N-1}f_{\ell,N}(y, \mathbf{z}^*) \quad \text{for } 2 \leq \ell \leq k \\ z_\ell &= hf_{\ell,2}(y, \mathbf{z}^*) + \dots + h^N f_{\ell,N+1}(y, \mathbf{z}^*) \quad \text{for } \ell > k \\ z_\ell &= 0 \quad \text{for } \ell \notin \mathcal{I}_N \end{aligned} \quad (3.17)$$

with initial values $z_\ell(0) = \mathcal{O}(h)$ for $2 \leq \ell \leq k$, the function

$$x(t) = y(t) + \sum_{\ell \in \mathcal{I}^*} \zeta_\ell^{t/h} z_\ell(t) \quad (3.18)$$

satisfies

$$\sum_{j=0}^k \alpha_j x(t+jh) = h \sum_{j=0}^k \beta_j f(x(t+jh)) + \mathcal{O}(h^{N+1}). \quad (3.19)$$

For $\mathbf{z}^* = 0$ the differential equation for y is the same as that of Theorem 3.1. The solutions of (3.17) satisfy $z_\ell(t) = \bar{z}_j(t)$ whenever $\zeta_\ell = \bar{\zeta}_j$ and this relation holds for the initial values. Moreover, $z_\ell(t) = \mathcal{O}(h^{m+1})$ on bounded time intervals if ζ_ℓ is a product of no fewer than $m \geq 2$ roots of $\rho(\zeta)$.

Proof. We let $z_1(t) := y(t)$ and insert the finite sum (3.18) into (3.19). This yields

$$\begin{aligned}
\sum_{j=0}^k \alpha_j x(t+jh) &= \sum_{j=0}^k \alpha_j \sum_{\ell \in \mathcal{I}} \zeta_\ell^{(t+jh)/h} e^{jhD} z_\ell(t) \\
&= \sum_{\ell \in \mathcal{I}} \zeta_\ell^{t/h} \sum_{j=0}^k \alpha_j \zeta_\ell^j e^{jhD} z_\ell(t) = \sum_{\ell \in \mathcal{I}} \zeta_\ell^{t/h} \rho(\zeta_\ell e^{hD}) z_\ell(t),
\end{aligned}$$

where, as usual, D represents differentiation with respect to time. We then expand $f(x(t))$ into a Taylor series around $y(t)$,

$$\begin{aligned}
f(x(t)) &= \sum_{m \geq 0} \frac{1}{m!} f^{(m)}(y(t)) \left(\sum_{\ell_1 \in \mathcal{I}^*} \zeta_{\ell_1}^{t/h} z_{\ell_1}(t), \dots, \sum_{\ell_m \in \mathcal{I}^*} \zeta_{\ell_m}^{t/h} z_{\ell_m}(t) \right) \\
&= \sum_{\ell \in \mathcal{I}} \zeta_\ell^{t/h} \sum_{m \geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1} \dots \zeta_{\ell_m} = \zeta_\ell} f^{(m)}(y(t)) (z_{\ell_1}(t), \dots, z_{\ell_m}(t)).
\end{aligned}$$

This gives, as above,

$$\begin{aligned}
&\sum_{j=0}^k \beta_j f(x(t+jh)) \\
&= \sum_{\ell \in \mathcal{I}} \zeta_\ell^{t/h} \sigma(\zeta_\ell e^{hD}) \sum_{m \geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1} \dots \zeta_{\ell_m} = \zeta_\ell} f^{(m)}(y(t)) (z_{\ell_1}(t), \dots, z_{\ell_m}(t)).
\end{aligned} \tag{3.20}$$

Comparing coefficients of $\zeta_\ell^{t/h}$ for $\ell \in \mathcal{I}_N$ in (3.19) thus yields

$$\rho(\zeta_\ell e^{hD}) z_\ell = h \sigma(\zeta_\ell e^{hD}) \sum_{m \geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1} \dots \zeta_{\ell_m} = \zeta_\ell} f^{(m)}(y) (z_{\ell_1}, \dots, z_{\ell_m}) \tag{3.21}$$

(for $\ell = 1$ and $m = 0$ the sum is understood to include the term $f(y)$). With the expansion $x \sigma(\zeta_\ell e^x) / \rho(\zeta_\ell e^x) = \mu_{\ell,0} + \mu_{\ell,1}x + \mu_{\ell,2}x^2 + \dots$ for $1 \leq \ell \leq k$, where ζ_ℓ is a simple root of $\rho(\zeta)$, this equation becomes

$$\dot{z}_\ell = \left(\mu_{\ell,0} + \mu_{\ell,1}hD + \dots \right) \sum_{m \geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1} \dots \zeta_{\ell_m} = \zeta_\ell} f^{(m)}(y) (z_{\ell_1}, \dots, z_{\ell_m}), \tag{3.22}$$

and with $\sigma(\zeta_\ell e^x) / \rho(\zeta_\ell e^x) = \mu_{\ell,0} + \mu_{\ell,1}x + \mu_{\ell,2}x^2 + \dots$ for $\ell > k$, where $\rho(\zeta_\ell) \neq 0$,

$$z_\ell = h \left(\mu_{\ell,0} + \mu_{\ell,1}hD + \dots \right) \sum_{m \geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1} \dots \zeta_{\ell_m} = \zeta_\ell} f^{(m)}(y) (z_{\ell_1}, \dots, z_{\ell_m}). \tag{3.23}$$

In the usual way (elimination of the first and higher derivatives by the differential equations and by the differentiated third relation of (3.17)) this allows us to define recursively the functions $f_{\ell,j}(y, \mathbf{z}^*)$.

From this construction process it follows that on bounded time intervals we have $z_\ell(t) = \mathcal{O}(h)$ for all $\ell \geq 2$, and $z_\ell(t) = \mathcal{O}(h^{m+1})$ if ζ_ℓ is a product of no fewer than $m \geq 2$ roots of $\rho(\zeta)$. In (3.20) and in the above construction of the coefficient functions $f_{\ell,j}(y, \mathbf{z}^*)$ we have neglected terms that contain at least N factors z_j . This gives rise to the $\mathcal{O}(h^{N+1})$ term in (3.19). \square

Initial values $y(0), z_\ell(0), \ell = 2, \dots, k$, for the system (3.17) are obtained from the starting approximations y_0, \dots, y_{k-1} via the relation

$$y_j = y(jh) + \sum_{\ell \in \mathcal{I}^*} \zeta_\ell^j z_\ell(jh), \quad j = 0, 1, \dots, k-1. \quad (3.24)$$

For $h = 0$ this represents a linear Vandermonde system for $y(0), z_\ell(0)$. The Implicit Function Theorem thus proves the local existence of a solution of (3.24) for sufficiently small step sizes h . If $y_j, j = 2, \dots, k$, approximate a solution $y_{ex}(t)$ of $\dot{y} = f(y)$ with an error $\mathcal{O}(h^s)$ (with $s \leq r+1$, where r is the order of the method), then $y(0) - y_{ex}(0) = \mathcal{O}(h^s)$ and $z_\ell(0) = \mathcal{O}(h^s)$ for $\ell = 2, \dots, k$.

The representation (3.16) of the numerical solution and the (principal and parasitic) modified equations (3.17) will be the main ingredients for the study of long-term stability of multistep methods in Sect. XV.5. An extension of the previous theorem to partitioned multistep methods is more or less straightforward. We leave the details as an exercise for the reader.

Multistep Methods for Second Order Differential Equations. A completely analogous result can be proved for stable, symmetric multistep methods (1.8) applied to $\ddot{y} = f(y)$. We again denote the zeros of $\rho(\zeta)$ by $\zeta_1 = 1$ and $\zeta_\ell, \ell = 2, \dots, q$. Notice, however, that $\zeta_1 = 1$ is always a double zero, and the others can be simple or double zeros of $\rho(\zeta)$, and that $q \leq k$. We consider the index sets $\mathcal{I}, \mathcal{I}^*, \mathcal{I}_N$, and \mathcal{I}_N^* as in (3.15).

Theorem 3.6. *Consider a stable, consistent, and symmetric multistep method (1.8). For every truncation index $N \geq 2$, there then exist h -independent functions $f_{\ell,j}(y, \dot{y}, \mathbf{z}^*)$ (where \mathbf{z}^* denotes the vector collecting as elements z_ℓ, \dot{z}_ℓ if ζ_ℓ is a double root, and z_ℓ if ζ_ℓ is a simple root of $\rho(\zeta)$) such that for every solution of*

$$\begin{aligned} \ddot{y} &= f_{1,1}(y, \dot{y}, \mathbf{z}^*) + hf_{1,2}(y, \dot{y}, \mathbf{z}^*) + \dots + h^{N-1}f_{1,N}(y, \dot{y}, \mathbf{z}^*) \\ \ddot{z}_\ell &= f_{\ell,1}(y, \dot{y}, \mathbf{z}^*) + \dots + h^{N-1}f_{\ell,N}(y, \dot{y}, \mathbf{z}^*) \quad \text{if } \rho(\zeta_\ell) = \rho'(\zeta_\ell) = 0 \\ \dot{z}_\ell &= hf_{\ell,2}(y, \dot{y}, \mathbf{z}^*) + \dots + h^Nf_{\ell,N+1}(y, \dot{y}, \mathbf{z}^*) \quad \text{if } \rho(\zeta_\ell) = 0, \rho'(\zeta_\ell) \neq 0 \\ z_\ell &= h^2f_{\ell,3}(y, \dot{y}, \mathbf{z}^*) + \dots + h^{N+1}f_{\ell,N+2}(y, \dot{y}, \mathbf{z}^*) \quad \text{if } \rho(\zeta_\ell) \neq 0 \\ z_\ell &= 0 \quad \text{for } \ell \notin \mathcal{I}_N \end{aligned} \quad (3.25)$$

with initial values $z_\ell(0) = \mathcal{O}(h)$ for $2 \leq \ell \leq q$, the function

$$x(t) = y(t) + \sum_{\ell \in \mathcal{I}^*} \zeta_\ell^{t/h} z_\ell(t) \quad (3.26)$$

satisfies

$$\sum_{j=0}^k \alpha_j x(t+jh) = h^2 \sum_{j=0}^k \beta_j f(x(t+jh)) + \mathcal{O}(h^{N+2}). \quad (3.27)$$

For $\mathbf{z}^* = 0$ the differential equation for y is the same as in (3.9). The solutions of (3.25) satisfy $z_\ell(t) = \bar{z}_j(t)$ whenever $\zeta_\ell = \bar{\zeta}_j$ and this relation holds for the initial values. Moreover, $z_\ell(t) = \mathcal{O}(h^{m+2})$ on bounded time intervals if ζ_ℓ is a product of no fewer than $m \geq 2$ roots of $\rho(\zeta)$.

Proof. In complete analogy to the proof of Theorem 3.5 we obtain

$$\rho(\zeta_\ell e^{hD})z_\ell = h^2 \sigma(\zeta_\ell e^{hD}) \sum_{m \geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1} \dots \zeta_{\ell_m} = \zeta_\ell} f^{(m)}(y)(z_{\ell_1}, \dots, z_{\ell_m}) \quad (3.28)$$

which differs from (3.21) only in the factor h^2 . Depending on whether ζ_ℓ is a double, a simple, or not a zero of $\rho(\zeta)$, we expand $x^2 \sigma(\zeta_\ell e^x)/\rho(\zeta_\ell e^x)$ or $x \sigma(\zeta_\ell e^x)/\rho(\zeta_\ell e^x)$ or $\sigma(\zeta_\ell e^x)/\rho(\zeta_\ell e^x)$ into a series of powers of x , and we denote its coefficients by $\mu_{\ell,j}$. This then yields

$$\ddot{z}_\ell = \left(\mu_{\ell,0} + \mu_{\ell,1} hD + \dots \right) \sum_{m \geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1} \dots \zeta_{\ell_m} = \zeta_\ell} f^{(m)}(y)(z_{\ell_1}, \dots, z_{\ell_m}), \quad (3.29)$$

if $\rho(\zeta_\ell) = \rho'(\zeta_\ell) = 0$, but $\rho''(\zeta_\ell) \neq 0$ (in particular for $\ell = 1$ and $\zeta_1 = 1$),

$$\dot{z}_\ell = h \left(\mu_{\ell,0} + \mu_{\ell,1} hD + \dots \right) \sum_{m \geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1} \dots \zeta_{\ell_m} = \zeta_\ell} f^{(m)}(y)(z_{\ell_1}, \dots, z_{\ell_m}), \quad (3.30)$$

if $\rho(\zeta_\ell) = 0$, but $\rho'(\zeta_\ell) \neq 0$, and

$$z_\ell = h^2 \left(\mu_{\ell,0} + \mu_{\ell,1} hD + \dots \right) \sum_{m \geq 0} \frac{1}{m!} \sum_{\zeta_{\ell_1} \dots \zeta_{\ell_m} = \zeta_\ell} f^{(m)}(y)(z_{\ell_1}, \dots, z_{\ell_m}), \quad (3.31)$$

if $\rho(\zeta_\ell) \neq 0$. The rest of the proof is identical to that of Theorem 3.5. \square

For the system of modified equations (3.25) we need initial values $y(0), \dot{y}(0), z_\ell(0), \dot{z}_\ell(0)$ if ζ_ℓ is a double root of $\rho(\zeta)$, and $z_\ell(0)$ if ζ_ℓ is a simple root. These initial values can be obtained from the starting approximations y_0, \dots, y_{k-1} via the relation (3.24).

Lemma 3.7. *Consider a stable, symmetric multistep method (1.8) of order r , and let the starting approximations y_0, \dots, y_{k-1} satisfy $y_j - y_{ex}(jh) = \mathcal{O}(h^s)$ with $2 \leq s \leq r+2$. Then there exist (locally) unique initial values for the system (3.25) such that its solution exactly satisfies (3.24).*

These initial values satisfy $z_\ell(0) = \bar{z}_j(0)$ if $\zeta_\ell = \bar{\zeta}_j$, and

$$\begin{aligned} y(0) - y_{ex}(0) &= \mathcal{O}(h^s), & h\dot{y}(0) - h\dot{y}_{ex}(0) &= \mathcal{O}(h^s), \\ z_\ell(0) &= \mathcal{O}(h^s), & h\dot{z}_\ell(0) &= \mathcal{O}(h^s) & \text{if } \zeta_\ell \text{ is a double root,} \\ z_\ell(0) &= \mathcal{O}(h^s), & & & \text{if } \zeta_\ell \text{ is a simple root.} \end{aligned} \quad (3.32)$$

Proof. We scale the derivatives by h , and consider $y(0), h\dot{y}(0), z_\ell(0)$ and $hz_\ell(0)$ as unknowns in the system (3.24), where $y(t)$ and $z_\ell(t)$ are a solution of (3.25). For $h = 0$ a linear, confluent Vandermonde system is obtained. Since this is an invertible matrix, the Implicit Function Theorem proves the statement. \square

XV.4 Can Multistep Methods be Symplectic?

Readers might be astonished to find a question mark in the title. The reason is that we shall present two definitions of symplecticity of multistep methods applied to a Hamiltonian system

$$\dot{p} = -H_q(p, q), \quad \dot{q} = H_p(p, q). \quad (4.1)$$

One works in the phase space of the exact flow, the other in a higher dimensional space. But which one is suitable? We further show that certain multistep methods can preserve energy over long times, even if they are not symplectic.

XV.4.1 Non-Symplecticity of the Underlying One-Step Method

A conjecture due to Feng Kang. (Y.-F. Tang 1993)

A natural definition of symplecticity consists of the requirement that the underlying one-step method (Theorem 2.4) be symplectic. This means that the (truncated) modified equation (3.1) is Hamiltonian. Unfortunately, we have the following negative result.

Theorem 4.1 (Tang 1993). *The underlying one-step method of a consistent linear multistep method (1.1) cannot be symplectic.*

Proof. We show that the first perturbation term in the modified equation (3.1) is in general not Hamiltonian. From (3.4) we know that $f_{r+1}(y) = \mu_r(D_1^r f)(y)$ which (omitting the non-zero error constant μ_r) is given by

$$\sum_{\tau \in T, |\tau|=r+1} \alpha(\tau) F(\tau)(y) = |\tau|! \sum_{\tau \in T, |\tau|=r+1} \frac{1}{\sigma(\tau)} b(\tau) F(\tau)(y) \quad (4.2)$$

with $b(\tau) = 1/\gamma(\tau)$ for $|\tau| = r+1$ (Theorem III.1.3 and (III.1.27)). Suppose now that (4.2) is Hamiltonian for all separable Hamiltonian vector fields $f(y) = J^{-1} \nabla H(y)$. Theorem IX.10.4 then implies

$$b(u \circ v) + b(v \circ u) = 0 \quad \text{for all } u, v \in T \text{ with } |u| + |v| = r+1.$$

This, however, is in contradiction with

$$\frac{1}{\gamma(u \circ v)} + \frac{1}{\gamma(v \circ u)} = \frac{1}{\gamma(u)} \cdot \frac{1}{\gamma(v)},$$

which is a consequence of Theorem VI.7.6 (because the exact solution is a symplectic transformation and, as a B-series, has coefficients $a(\tau) = 1/\gamma(\tau)$). \square

A similar negative result holds for a much larger class of integration methods. For example, it is proved by Hairer & Leone (1998) that, among the class of one-leg methods (see (4.7) below), only the implicit mid-point rule is symplectic (in the sense that the underlying one-step method is symplectic).

Partitioned Linear Multistep Methods. We know at least one symplectic method of the form (1.14). It is the symplectic Euler method (VI.3.1), which combines the implicit and the explicit Euler methods. However, we do not have better within the class of partitioned multistep methods as is shown in the next theorem.

Theorem 4.2. *If the underlying one-step method of a consistent, partitioned linear multistep method (1.14) is symplectic for all separable Hamiltonian systems, then its order satisfies $r \leq 1$.*

Proof. Suppose that the order of the method is $r \geq 2$. By (3.8), the dominant perturbation term in the modified differential equation is $\mu_r^{(A)} h^r (D_1^r f)(y, z)$ for the y -component and $\mu_r^{(B)} h^r (D_1^r g)(y, z)$ for the z -component (at least one of the coefficients $\mu_r^{(A)}$ and $\mu_r^{(B)}$ is non-zero). This is a P-series with coefficients $b(\tau) = \mu_r^{(A)} / \gamma(\tau)$ if $\tau \in TP_p, |\tau| = r + 1$ and $b(\tau) = \mu_r^{(B)} / \gamma(\tau)$ if $\tau \in TP_q, |\tau| = r + 1$. If the underlying one-step method is symplectic (i.e., the modified differential equation is locally Hamiltonian), Theorem IX.10.4 implies that

$$b(u \circ v) + b(v \circ u) = 0 \quad \text{for } u \in TP_p, v \in TP_q, |u| + |v| = r + 1. \quad (4.3)$$

Taking for $u \in TP_p$ the tree with one vertex, and for $v \in TP_q$ an arbitrary tree with $|v| = r$, condition (4.3) gives the first relation of

$$\frac{\mu_r^{(A)}}{(r+1)\gamma(v)} + \frac{\mu_r^{(B)} r}{(r+1)\gamma(v)} = 0, \quad \frac{\mu_r^{(B)}}{(r+1)\gamma(v)} + \frac{\mu_r^{(A)} r}{(r+1)\gamma(v)} = 0.$$

Exchanging the colour of the vertices gives the second relation. This contradicts our assumption $r \geq 2$. \square

If we restrict our considerations to Hamiltonian systems with

$$H(p, q) = \frac{1}{2} p^T C p + c^T p + U(q), \quad (4.4)$$

where the kinetic energy is at most quadratic in p , we can find symplectic, partitioned multistep methods of order two. Indeed, the combination of the trapezoidal rule with the explicit midpoint rule

$$p_{n+1} - p_n = -\frac{h}{2} \left(\nabla U(q_{n+1}) + \nabla U(q_n) \right), \quad q_{n+1} - q_{n-1} = 2h(Cp_n + c) \quad (4.5)$$

has the Störmer–Verlet method as underlying one-step method. This is seen as follows: since the Hamiltonian is separable, formula (VI.3.4) yields the first formula of (4.5). The second relation is a consequence of $q_{n+1} - q_n + h(Cp_{n+1/2} + c)$ and $p_{n+1/2} + p_{n-1/2} = 2p_n$, and uses the linearity of $H_p(p, q)$.

Also for this special class of Hamiltonian systems we cannot achieve high order and symplecticity at the same time.

Theorem 4.3. *If the underlying one-step method of a consistent, partitioned linear multistep method (1.14) is symplectic for all Hamiltonian systems with Hamiltonian of the form (4.4), then its order satisfies $r \leq 2$.*

Proof. The beginning is the same as that for Theorem 4.2. We let $r \geq 2$ be the order of the method (A) so that $\mu_r^{(A)} \neq 0$. Instead of (4.3) we now have to use

$$b(u \circ \circ v) - b(v \circ \circ u) = 0 \quad \text{for } u, v \in TN_p, |u| + |v| = r, \quad (4.6)$$

which also follows from Theorem IX.10.4. Taking for $u \in TN_p$ the tree with one vertex, and for $v \in TN_p$ an arbitrary tree with $|v| = r - 1$, condition (4.6) gives the relation

$$\frac{\mu_r^{(A)}(r-1)}{2(r+1)\gamma(v)} - \frac{\mu_r^{(A)}}{r(r+1)\gamma(v)} = 0,$$

which is contradictory for $r > 2$, because $\mu_r^{(A)} \neq 0$. \square

Remark 4.4. We believe that the statement of Theorem 4.3 remains true, if we restrict our consideration to Hamiltonian functions (4.4) with $c = 0$ and invertible matrix C . Since multistep methods (1.8) for second order differential equations can be converted into partitioned multistep methods, this then implies that methods (1.8) cannot be symplectic unless the order satisfies $r \leq 2$.

XV.4.2 Symplecticity in the Higher-Dimensional Phase Space

We present here a second approach for the definition of symplecticity of multistep methods (more precisely, of one-leg methods). It is much inspired by the G -stability theory of Dahlquist (1975) for the study of stiff differential equations.

To simplify the nonlinear stability theory of linear multistep methods (1.1), Dahlquist (1975) introduced the so-called *one-leg methods*, which are defined by the relation

$$\sum_{j=0}^k \alpha_j y_{n+j} = hf \left(\sum_{j=0}^k \beta_j y_{n+j} \right), \quad (4.7)$$

where the normalization $\sigma(1) = \sum_j \beta_j = 1$ is assumed. In fact, there is a close relationship between the numerical solution of (4.7) and (1.1), and their long-time behaviour is the same (see Sect. V.6 of Hairer & Wanner, 1996). In the following we consider the super-vectors $Y_n = (y_{n+k-1}, \dots, y_n)^T$ collecting k consecutive approximations of the solution.

Definition 4.5. Let G be an invertible symmetric matrix of dimension k . A k -step multistep or one-leg method is called *G -symplectic* if

$$Y_{n+1}^T (G \otimes S) Y_{n+1} = Y_n^T (G \otimes S) Y_n, \quad (4.8)$$

whenever the differential equation $\dot{y} = f(y)$ has $y^T S y$ as invariant (with symmetric S), i.e., the vector field satisfies $y^T S f(y) = 0$ for all y .

It is of course also possible to express this definition in terms of differential forms. As a consequence of Lemma VI.4.1 the conservation of quadratic first integrals is equivalent to symplecticity (Bochev & Scovel 1994).

In contrast to the negative results of Sect. XV.4.1, there exist a lot of G -symplectic methods. We have the following result.

Theorem 4.6 (Eirola & Sanz-Serna 1992). *Every irreducible symmetric one-leg method (4.7) is G -symplectic for some matrix G .*

Proof. We recall that a one-leg method is irreducible if the generating polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ have no common zeros.

Construction of G . The symmetry relation (1.3) implies $\rho(1/\zeta) = -\zeta^{-k}\rho(\zeta)$ and $\sigma(1/\zeta) = \zeta^{-k}\sigma(\zeta)$. Consequently, the polynomial $\rho(\zeta)\sigma(\omega) + \rho(\omega)\sigma(\zeta)$ vanishes for $\omega = 1/\zeta$, and contains the factor $\zeta\omega - 1$. We then define G by

$$\frac{1}{2}(\rho(\zeta)\sigma(\omega) + \rho(\omega)\sigma(\zeta)) = (\zeta\omega - 1) \sum_{i,j=1}^k g_{ij} \zeta^{i-1} \omega^{j-1}. \quad (4.9)$$

The matrix G obtained in this way is symmetric.

Regularity of G . Applying the geometric series we get

$$\sum_{i,j=1}^k g_{ij} \zeta^{i-1} \omega^{j-1} = -\frac{1}{2}(\rho(\zeta)\sigma(\omega) + \rho(\omega)\sigma(\zeta))(1 + \zeta\omega + \zeta^2\omega^2 + \dots),$$

where the identity holds as formal power series. Suppose that the matrix G is not invertible. Then there exists a vector $u = (u_0, u_1, \dots, u_{k-1})^T$ such that $Gu = 0$. We formally replace the appearances of ω^{j-1} with u_{j-1} for $j \leq k$ and with zero for $j > k$. This gives an identity of the form $0 = \rho(\zeta)a(\zeta) + \sigma(\zeta)b(\zeta)$ with polynomials $a(\zeta)$ and $b(\zeta)$ of degree at most $k-1$, and we get a contradiction with the irreducibility of the method.

G -Symplecticity. We next replace in (4.9) $\zeta^i \omega^j$ with $y_{n+i}^T S y_{n+j}$. Together with (4.7) this yields

$$h \left(\sum_{i=0}^k \beta_i y_{n+i} \right)^T S f \left(\sum_{i=0}^k \beta_i y_{n+i} \right) = Y_{n+1}^T (G \otimes S) Y_{n+1} - Y_n^T (G \otimes S) Y_n,$$

where $Y_n = (y_n, \dots, y_{n+k-1})^T$. This proves (4.8) for all functions $f(y)$ satisfying $y^T S f(y) = 0$. \square

Example 4.7. We consider the explicit midpoint rule (1.6), which is also a one-leg method, and the 3-step method (3.13). By Theorem 4.6 the one-leg versions are G -symplectic. Following the constructive proof of this theorem we find

$$G = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad G = \begin{pmatrix} 0 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & 0 \end{pmatrix},$$

respectively. We apply both methods to two closely related Hamiltonian systems, namely the pendulum equation with $H(p, q) = p^2/2 - \cos q$ and a perturbed problem with $H(p, q) = p^2/2 - \cos q(1 - p/6)$, and we study the preservation of the

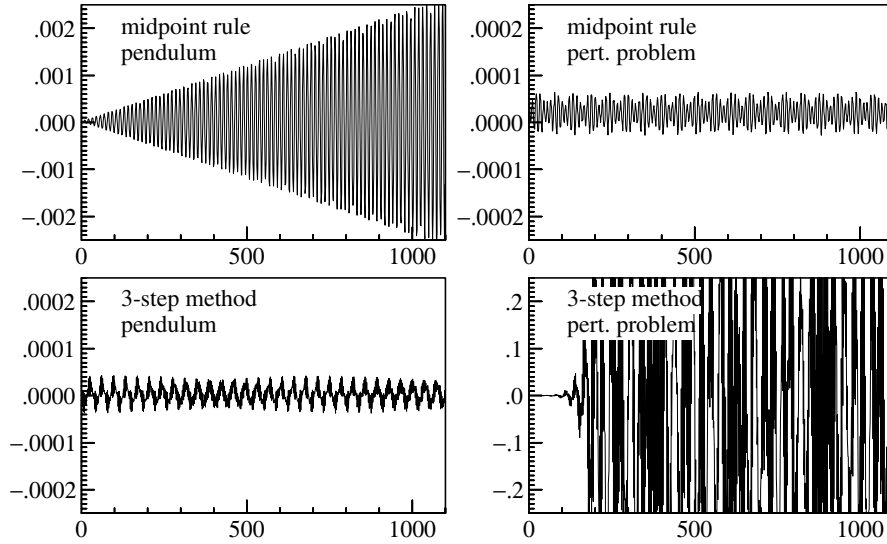


Fig. 4.1. Numerical Hamiltonian $H(p_n, q_n) - H(p_0, q_0)$ of the explicit mid-point rule and the 3-step method (3.13), applied with step size $h = 0.01$ to the pendulum problem ($H(p, q) = p^2/2 - \cos q$) and to a perturbed problem ($H(p, q) = p^2/2 - \cos q(1 - p/6)$) on the interval $[0, 1100]$ (only every 131st step is drawn)

Hamiltonian (see Fig. 4.1). The result is somewhat surprising. The midpoint rule behaves well for the perturbed problem, but shows a linear error growth in the Hamiltonian for the pendulum problem. On the other side, the weakly stable 3-step method behaves well for the pendulum equation (which is in agreement with the stable behaviour of Fig. 3.1), but has an exponential error growth for the perturbed problem. Notice that different scales are used in the four pictures.

The above example illustrates that G -symplecticity of a numerical method is not sufficient for a good long-time behaviour. It is necessary to get under control the parasitic solution components.

XV.4.3 Modified Hamiltonian of Multistep Methods

After the negative results of Sect. XV.4.1, we are fortunately also able to prove positive results concerning the near conservation of the Hamiltonian.

Theorem 4.8. *For a symmetric, consistent linear multistep method (1.1) of order r applied to $\dot{y} = J^{-1}\nabla H(y)$, there exists a series of the form*

$$\tilde{H}(y) = H(y) + h^r H_{r+1}(y) + h^{r+2} H_{r+3}(y) + \dots, \quad (4.10)$$

which is a formal first integral of the modified equation (3.1) without truncation.

Proof. With $\rho(e^x)/(x\sigma(e^x)) = 1 + \gamma_r x^r + \gamma_{r+2} x^{r+2} + \dots$ it follows from (3.3) that the solution of the modified differential equation satisfies

$$(1 + \gamma_r h^r D^r + \gamma_{r+2} h^{r+2} D^{r+2} + \dots) \dot{y} = J^{-1} \nabla H(y) + \mathcal{O}(h^N), \quad (4.11)$$

where, due to the symmetry of the method, only odd derivatives of $y(t)$ appear. We multiply both sides with $\dot{y}^T J$ so that the right-hand side becomes the total derivative $\frac{d}{dt} H(y)$. On the left-hand side we note $\dot{y}^T J \dot{y} = 0$, $\dot{y}^T J y^{(3)} = \frac{d}{dt} (\dot{y}^T J \ddot{y})$ and similarly for higher derivatives

$$\dot{y}^T J y^{(2m+1)} = \frac{d}{dt} (\dot{y}^T J y^{(2m)} - \ddot{y}^T J y^{(2m-1)} + \dots \pm y^{(m)T} J y^{(m+1)}). \quad (4.12)$$

We thus obtain a time derivative of an expression in which the appearing derivatives can be substituted as functions of y via the modified differential equation (3.1). Altogether this yields

$$-\frac{d}{dt} (h^r H_{r+1}(y) + h^{r+2} H_{r+3}(y) + \dots) = \frac{d}{dt} H(y) + \mathcal{O}(h^N).$$

which proves the statement. \square

The statement of the previous theorem is somewhat surprising. The underlying one-step method, although not symplectic, nearly conserves the Hamiltonian for general $H(y)$ (not even reversibility is required). This indicates that the condition (IX.9.20) can be satisfied for all trees also by non-symplectic methods.

For partitioned multistep methods we do not know of a similar result unless if we restrict our consideration to Hamiltonians of the form (4.4). In this case we are concerned with multistep methods for second order differential equations.

Theorem 4.9. *For a symmetric, consistent linear multistep method (1.8) of order r applied to $\ddot{y} = -\nabla U(y)$, there exists a series of the form*

$$\tilde{H}(y, \dot{y}) = \frac{1}{2} \dot{y}^T \dot{y} + U(y) + h^r H_{r+1}(y, \dot{y}) + h^{r+2} H_{r+3}(y, \dot{y}) + \dots, \quad (4.13)$$

which is a formal first integral of the modified equation (3.9) without truncation.

Proof. The proof is very similar to that of the previous theorem. We expand $\rho(e^x)/(x^2 \sigma(e^x)) = 1 + \gamma_r x^r + \gamma_{r+2} x^{r+2} + \dots$, and similar to (3.10) we obtain

$$(1 + \gamma_r h^r D^r + \gamma_{r+2} h^{r+2} D^{r+2} + \dots) \ddot{y} = -\nabla U(y) + \mathcal{O}(h^N). \quad (4.14)$$

This time we multiply both sides with \dot{y}^T . The right-hand side becomes the total derivative $\frac{d}{dt} U(y)$, and for the left-hand side we use $\dot{y}^T \ddot{y} = \frac{d}{dt} (\dot{y}^T \dot{y})$ and for higher even-order derivatives

$$\dot{y}^T y^{(2m)} = \frac{d}{dt} (\dot{y}^T y^{(2m-1)} - \ddot{y}^T y^{(2m-2)} + \dots \pm \frac{1}{2} y^{(m)T} y^{(m)}). \quad (4.15)$$

Integrating and substituting second and higher derivatives of y via the modified differential equation (3.9) yields the desired formal first integral close to the Hamiltonian of the system. \square

The formal first integral (4.13) does not depend on how approximations to the derivative $v = \dot{y}$ are obtained. If the derivative at grid points is numerically computed with the formula (3.11), then one can use the one-to-one correspondence (3.12) to express the coefficient functions of the modified differential equation in terms of y and v .

XV.4.4 Modified Quadratic First Integrals

Symplectic one-step methods exactly preserve quadratic first integrals (Sect. IV.2). This is not true for the underlying one-step method of symmetric multistep methods. However, as we shall prove in this section, it nearly preserves such first integrals.

Theorem 4.10. *Let $Q(y) = y^T C y$ (with a symmetric matrix C) be a first integral of $\dot{y} = f(y)$. For a symmetric, consistent linear multistep method (1.1) of order r , there then exists a series of the form*

$$\tilde{Q}(y) = y^T C y + h^r Q_{r+1}(y) + h^{r+2} Q_{r+3}(y) + \dots, \quad (4.16)$$

which is a formal first integral of the modified equation (3.1) without truncation.

Proof. We multiply (4.11) with $y^T C$ and thus obtain

$$y^T C (1 + \gamma_r h^r D^r + \gamma_{r+2} h^{r+2} D^{r+2} + \dots) \dot{y} = y^T C f(y) + \mathcal{O}(h^N).$$

Since $y^T C y$ is a first integral, the term on the right-hand side vanishes. For the terms on the left-hand side we notice that $y^T C \dot{y} = \frac{1}{2} \frac{d}{dt} (y^T C y)$ and that

$$y^T C y^{(2m+1)} = \frac{d}{dt} \left(y^T C y^{(2m)} - \dot{y}^T C y^{(2m-1)} + \dots \pm \frac{1}{2} y^{(m)T} C y^{(m)} \right). \quad (4.17)$$

As in the proofs of Sect. XV.4.3 we now deduce the statement. \square

A similar result holds for second order differential equations and methods (1.8). This concerns for example the total angular momentum in N -body systems.

Theorem 4.11. *Suppose that $\ddot{y} = f(y)$ has $L(y, \dot{y}) = y^T E \dot{y}$ as first integral, i.e., E is skew-symmetric and $y^T E f(y) = 0$. For a symmetric, consistent linear multistep method (1.8) of order r , there then exists a series of the form*

$$\tilde{L}(y, \dot{y}) = y^T E \dot{y} + h^r L_{r+1}(y, \dot{y}) + h^{r+2} L_{r+3}(y, \dot{y}) + \dots, \quad (4.18)$$

which is a formal first integral of the modified equation (3.9) without truncation.

Proof. Multiplying (4.14) with $y^T E$ gives

$$y^T E (1 + \gamma_r h^r D^r + \gamma_{r+2} h^{r+2} D^{r+2} + \dots) \ddot{y} = y^T E f(y) + \mathcal{O}(h^N).$$

The term at the right vanishes. Since E is a skew-symmetric matrix, we have for the terms to the left that $y^T E \ddot{y} = \frac{d}{dt} y^T E \dot{y}$ and that

$$y^T E y^{(2m+2)} = \frac{d}{dt} \left(y^T E y^{(2m+1)} - \dot{y}^T E y^{(2m)} + \dots \pm y^{(m)T} E y^{(m+1)} \right). \quad (4.19)$$

This yields the statement as in the previous proofs. \square

Remark 4.12. Noticing that the underlying one-step method of a symmetric multistep method can be expressed as a formal B-series (cf. Sect. XV.2.2), it follows from (4.17) that the modified first integral of Theorem 4.10 is of the form (VI.8.6). By Theorem VI.8.5 the underlying one-step method is therefore conjugate to a symplectic integrator.

A similar result holds for symmetric methods (1.8) complemented with a symmetric derivative approximation (3.11). The variables v and \dot{y} are related via (3.12) having an expansion in even powers of h . Substituting $\dot{y} = \dot{y}(y, v)$ of this relation into the modified first integral (4.18), we obtain an expression of the form (VI.8.11). Here, the elementary differentials correspond to the system $\dot{y} = v$, $\dot{v} = f(y)$ (v has to be identified with z). Theorem VI.8.8 combined with Theorem 4.11 proves that the underlying one-step method is conjugate to a symplectic integrator.

XV.5 Long-Term Stability

The results of Sects. XV.4.3 and XV.4.4 imply the near conservation of the total energy and of the angular momentum in N -body problems for numerical solutions of the underlying one-step method of multistep methods. This, however, is of no value as long as the parasitic solutions of the multistep method are not under control. The present section is devoted to the study of the stability of numerical solutions over long time intervals.

XV.5.1 Role of Growth Parameters

The analysis of this section is based on the representation

$$y_n = y(nh) + \sum_{\ell \in \mathcal{I}^*} \zeta_\ell^n z_\ell(nh) \quad (5.1)$$

of the numerical solution of a multistep method (cf. formula (3.16)).

Linear Multistep Methods for First Order Equations. By Theorem 3.5 the parasitic components z_ℓ (for $2 \leq \ell \leq k$) are the solution of a differential equation which, by (3.22), is of the form

$$\dot{z}_\ell = \mu_\ell f'(y) z_\ell + \dots \quad (5.2)$$

This is just the variational equation of $\dot{y} = f(y)$ scaled by

$$\mu_\ell = \frac{\sigma(\zeta_\ell)}{\zeta_\ell \rho'(\zeta_\ell)}, \quad (5.3)$$

which is the so-called *growth parameter* as introduced by Dahlquist (1959) and motivated there by a linear stability analysis (see Exercise 5).

We shall illustrate at the examples of Sect. XV.3.2 that the study of the truncated equation (5.2) gives already a lot of insight into the long-time behaviour of multistep methods.

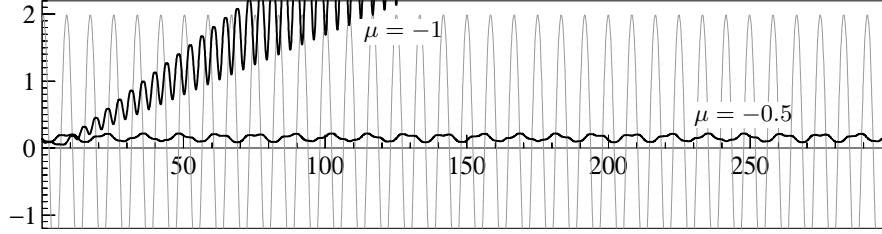


Fig. 5.1. First component of the solution of the pendulum equation (grey) together with the Euclidean norm of the solution $v(t)$ of the scaled variational equation (5.4)

Example 5.1. For the pendulum equation, the truncated equation (5.2) is

$$\begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{pmatrix} y_2 \\ -\sin y_1 \end{pmatrix}, \quad \begin{pmatrix} \dot{v}_1 \\ \dot{v}_2 \end{pmatrix} = \mu \begin{pmatrix} 0 & 1 \\ -\cos y_1 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}. \quad (5.4)$$

We fix initial values as $y(0) = (1.9, 0.4)^T$ and $v(0) = (0.1, 0.1)^T$. Figure 5.1 shows the solution component $y_1(t)$ in grey, and the Euclidean norm of $v(t)$ as solid black line, once for $\mu = -1$ and once for $\mu = 0.5$. We notice that the function $v(t)$ remains small and bounded for $\mu = -0.5$, and that it increases linearly for $\mu = -1$.

This agrees perfectly with the observations of Figs. 3.1 and 3.2, because the method (3.13) has growth parameter $\mu_\ell = -0.5$ for the roots $\zeta_\ell = \pm i$, whereas the explicit midpoint rule (3.14) has $\mu_\ell = -1$ for $\zeta_\ell = -1$.

The same analysis for partitioned multistep methods allows one to better understand the behaviour of the different methods in Fig. 1.3. The leading term of the parasitic modified equations depends on whether ζ_ℓ is a root of both polynomials $\rho_A(\zeta)$ and $\rho_B(\zeta)$, or only of one of them. This is very similar to the situation encountered with multistep methods for second order differential equations which we treat next.

Linear Multistep Methods for Second Order Equations. Theorem 3.6 tells us that the modified equation for the parasitic components z_ℓ depends on the multiplicity of the root ζ_ℓ . Consider a stable, symmetric method (1.8) for $\ddot{y} = f(y)$. If ζ_ℓ is a double root of $\rho(\zeta)$, formula (3.29) yields

$$\ddot{z}_\ell = \mu_\ell f'(y) z_\ell + \dots, \quad \mu_\ell = \frac{2\sigma(\zeta_\ell)}{\zeta_\ell^2 \rho''(\zeta_\ell)}, \quad (5.5)$$

where we have not written terms containing at least two factors z_j . If ζ_ℓ is a single root of $\rho(\zeta)$, we get from (3.30) that

$$\dot{z}_\ell = h\mu_\ell f'(y) z_\ell + \dots, \quad \mu_\ell = \frac{\sigma(\zeta_\ell)}{\zeta_\ell \rho'(\zeta_\ell)}. \quad (5.6)$$

There is an enormous difference between the parasitic modified equations corresponding to double or single roots of $\rho(\zeta)$. Equation (5.5) is the complete analogue

of (5.2) and, as before, the long-time behaviour is hardly predictable and strongly depends on the growth parameter. For single roots, however, we are concerned with a first order differential equation (5.6) having an additional factor h as bonus. For the analysis of Sects. XV.5.2 and XV.5.3 it is important to have only single roots.

Definition 5.2. A symmetric multistep method (1.8) for second order differential equations is called *s-stable* if, apart from the double root at 1, all zeros of $\rho(\zeta)$ are simple and of modulus one.

The linearized parasitic modified equations give much insight into the long-time behaviour of multistep methods. To get rigorous estimates over long times, however, further considerations are necessary. A partial result is given by Cano & Sanz-Serna (1998) for multistep methods (1.8) applied to equations $\ddot{y} = f(y)$ with periodic exact solution. There, the first terms of the asymptotic error expansion for the global error are computed, and their growth as a function of time is studied. We shall follow the approach of Hairer & Lubich (2004) who exploit the Hamiltonian structure of second order differential equations.

XV.5.2 Hamiltonian of the Full Modified System

In the remainder of this section we restrict our consideration to s-stable, irreducible linear multistep methods

$$\sum_{j=0}^k \alpha_j q_{n+j} = -h^2 \sum_{j=0}^k \beta_j \nabla U(q_{n+j}), \quad (5.7)$$

applied to Hamiltonian systems written as

$$\ddot{q} = -\nabla U(q), \quad (5.8)$$

where $U(q)$ is assumed to be real-analytic in the considered region.

The key to proving long-time error estimates is the observation that much of the Hamiltonian structure is conserved in the modified equations (3.25). The results and techniques of this subsection are closely related to those of Sect. XIII.6.3 developed for numerical methods for oscillatory differential equations.

We let $\mathbf{z} = (z_\ell)_{\ell \in \mathcal{I}_N}$ and define $\mathcal{U}(\mathbf{z})$ as

$$\mathcal{U}(\mathbf{z}) = U(z_0) + \sum_{m \geq 1} \frac{1}{m!} \sum_{\zeta_{\ell_1} \dots \zeta_{\ell_m} = 1} U^{(m)}(z_0)(z_{\ell_1}, \dots, z_{\ell_m}), \quad (5.9)$$

where the second sum is over all indices $\ell_1 \in \mathcal{I}_N^*, \dots, \ell_m \in \mathcal{I}_N^*$ with $\zeta_{\ell_1} \dots \zeta_{\ell_m} = 1$ (using the notation of Sect. XV.3.2). Since the roots of $\rho(\zeta)$, different from $\zeta_1 = 1$ are complex and appear in pairs (Exercise 3), also the functions z_ℓ appear in pairs. It is convenient to use the notation $z_{-\ell} = z_j$ if $\bar{\zeta}_\ell = \zeta_j$.

It follows from (3.28) with $f(q) = -\nabla U(q)$ that every solution of the truncated modified equation (3.25) satisfies

$$\rho(\zeta_\ell e^{hD})z_\ell = -h^2\sigma(\zeta_\ell e^{hD})\nabla_{z_\ell}\mathcal{U}(\mathbf{z}) + \mathcal{O}(h^{N+2}) \quad (5.10)$$

(for all $\ell \in \mathcal{I}$) as long as

$$y \in K, \quad \|\dot{y}\| \leq M, \quad \|z_\ell\| \leq \delta \text{ for } 1 < \ell < k, \quad (5.11)$$

where K is a compact subset of the domain of analyticity of $U(q)$, $M > 0$ some bound on the derivative, and $0 < \delta = \mathcal{O}(h)$ is a sufficiently small constant (note that this implies $\|z_\ell\| \leq \delta$ for all $\ell \in \mathcal{I}^*$ if h is sufficiently small, cf. the algebraic relations of (3.25)).

For ease of presentation, we assume for the moment that $\sigma(\zeta_\ell) \neq 0$ for all $\ell \in \mathcal{I}_N$ (we know that this holds for $1 \leq \ell < k$, because the method is irreducible). We apply the operator $\sigma^{-1}(\zeta_\ell e^{hD})$ to both sides of (5.10) and divide by h^2 :

$$h^{-2}\left(\frac{\rho}{\sigma}\right)(\zeta_\ell e^{hD})z_\ell = -\nabla_{z_\ell}\mathcal{U}(\mathbf{z}) + \mathcal{O}(h^N). \quad (5.12)$$

We then multiply with $\dot{z}_{-\ell}^T$, sum over all $\ell \in \mathcal{I}_N$, and thus obtain

$$h^{-2} \sum_{\ell \in \mathcal{I}_N} \dot{z}_{-\ell}^T \left(\frac{\rho}{\sigma}\right)(\zeta_\ell e^{hD})z_\ell + \frac{d}{dt}\mathcal{U}(\mathbf{z}) = \mathcal{O}(h^N). \quad (5.13)$$

We now show that also the first expression on the left-hand side is a total derivative of a function depending on \mathbf{z} and its time derivatives. For this we note that

$$\left(\frac{\rho}{\sigma}\right)(\zeta_\ell e^{ix}) = \sum_{j \geq 0} c_{\ell,j} x^j \quad \text{with real coefficients} \quad c_{\ell,j} = (-1)^j c_{-\ell,j}. \quad (5.14)$$

This holds because the symmetry of the multistep method yields $(\rho/\sigma)(1/\zeta) = (\rho/\sigma)(\zeta)$ and hence, for real x , $(\rho/\sigma)(\zeta_\ell e^{ix}) = (\rho/\sigma)(\overline{\zeta_\ell e^{ix}}) = (\rho/\sigma)(\zeta_{-\ell} e^{ix})$. With the expansion (5.14) we obtain

$$\left(\frac{\rho}{\sigma}\right)(\zeta_\ell e^{hD})z_\ell = \sum_{j=0}^{N+1} c_{\ell,j} (-ih)^j \dot{z}_\ell^{(j)} + \mathcal{O}(h^{N+2}). \quad (5.15)$$

To study (5.13) we apply the relation (4.12) for the real function $y = z_1$ and for z_ℓ corresponding to $\zeta_\ell = -1$, while for the complex-valued functions $z = z_\ell$, with complex conjugate $\bar{z} = z_{-\ell}$, we use

$$\begin{aligned} \operatorname{Re} \dot{\bar{z}}^T z^{(2m)} &= \operatorname{Re} \frac{d}{dt} \left(\dot{\bar{z}}^T z^{(2m-1)} - \ddot{\bar{z}}^T z^{(2m-2)} + \dots \pm \frac{1}{2} (\bar{z}^{(m)})^T z^{(m)} \right) \\ \operatorname{Im} \dot{\bar{z}}^T z^{(2m+1)} &= \operatorname{Im} \frac{d}{dt} \left(\dot{\bar{z}}^T z^{(2m)} - \ddot{\bar{z}}^T z^{(2m-1)} + \dots \mp (\bar{z}^{(m)})^T z^{(m+1)} \right). \end{aligned}$$

Together with (5.15) these relations show that the terms

$$\begin{aligned} \dot{z}_{-\ell}^T \left(\frac{\rho}{\sigma}\right)(\zeta_\ell e^{hD})z_\ell + \dot{z}_\ell^T \left(\frac{\rho}{\sigma}\right)(\zeta_{-\ell} e^{hD})z_{-\ell} \\ = \sum_{j=0}^{N+1} c_{\ell,j} 2 \operatorname{Re} \left((-ih)^j \dot{\bar{z}}_\ell^T z_\ell^{(j)} \right) + \mathcal{O}(h^{N+2}) \end{aligned}$$

give a total derivative (up to the remainder term). Hence the left-hand side of (5.13) can be written as the time derivative of a function which depends on z_ℓ , $\ell \in \mathcal{I}_N$, and on their derivatives. Using the modified equation (3.25) we eliminate all z_ℓ corresponding to ζ_ℓ with $\rho(\zeta_\ell) \neq 0$ and their derivatives, the first and higher derivatives of z_ℓ (for $1 < \ell < k$), and the second and higher derivatives of $y = z_1$. We thus get a function

$$\mathcal{H}(y, \dot{y}, \mathbf{z}^*) = H_0(y, \dot{y}, \mathbf{z}^*) + \dots + h^{N-1} H_{N-1}(y, \dot{y}, \mathbf{z}^*) \quad (5.16)$$

with $\mathbf{z}^* = (z_\ell)_{\ell=2}^{k-1}$, such that

$$\frac{d}{dt} \mathcal{H}(y(t), \dot{y}(t), \mathbf{z}^*(t)) = \mathcal{O}(h^N), \quad (5.17)$$

along solutions of (3.25) that stay in a set defined by (5.11). The function \mathcal{H} is therefore an almost-invariant of the system (3.25).

If, however, $\sigma(\zeta)$ does have a zero ζ_ℓ , then we omit the corresponding term from the sum in (5.13). Hence the term $\dot{z}_{-\ell}^T \nabla_{z_{-\ell}} \mathcal{U}(\mathbf{z})$ is missing from $(d/dt)\mathcal{U}(\mathbf{z})$ and must therefore be compensated in the remainder term. Since ζ_ℓ is a product of no fewer than two zeros of $\rho(\zeta)$, it follows from (3.31) and from $\mu_{\ell,0} = 0$ that $z_\ell = \mathcal{O}(h^3 \delta^2)$, as long as $\|z_j\| \leq \delta$ for $1 < j < k$. We further have $\nabla_{z_{-\ell}} \mathcal{U}(\mathbf{z}) = \mathcal{O}(\delta^2)$, so that the remainder term in (5.17) is augmented by $\mathcal{O}(h^3 \delta^4)$.

We summarize the above considerations (Hairer & Lubich 2004) as follows.

Theorem 5.3. *Every solution of the truncated modified equation (3.25) satisfies, with \mathcal{H} from (5.16),*

$$\mathcal{H}(y(t), \dot{y}(t), \mathbf{z}^*(t)) = \mathcal{H}(y(0), \dot{y}(0), \mathbf{z}^*(0)) + \mathcal{O}(th^N) + \mathcal{O}(th^3 \delta^4) \quad (5.18)$$

as long as the solution stays in the set defined by (5.11). Moreover,

$$\mathcal{H}(y, \dot{y}, \mathbf{z}^*) = H(y, \dot{y}) + \mathcal{O}(h^p) + \mathcal{O}(h\delta^2). \quad (5.19)$$

The closeness to the Hamiltonian $H(y, \dot{y}) = \frac{1}{2} \|\dot{y}\|^2 + U(y)$ follows also directly from the above construction. For $\mathbf{z}^* = 0$ we have $\mathcal{H}(y, \dot{y}, 0) = \tilde{H}(y, \dot{y})$, where \tilde{H} is the modified energy from Theorem 4.9.

We will use Theorem 5.3 in Section XV.6.1 to infer the long-time near-conservation of the Hamiltonian along numerical solutions. Before that we need to bound the parasitic components.

XV.5.3 Long-Time Bounds for Parasitic Solution Components

The modified equations have further almost-invariants which are close to the squares of the norms of the parasitic components that correspond to the roots of $\rho(\zeta)$. We derive them here and use them to show that all parasitic solution components remain small over very long times. The techniques used in this subsection are similar to those in Sects. XIII.6 and XIII.7.

We consider ℓ with $1 < \ell < k$ for which ζ_ℓ is a *simple* root of $\rho(\zeta)$ and $\sigma(\zeta_\ell) \neq 0$. The dominant term on the left-hand side of (5.12) is $-c_{\ell,1}ih^{-1}\dot{z}_\ell$. Since

$$\frac{d}{dt}\|z_\ell\|^2 = z_{-\ell}^T \dot{z}_\ell + z_\ell^T \dot{z}_{-\ell}, \quad (5.20)$$

we multiply (5.12) with $z_{-\ell}^T$ and the corresponding equation for $\zeta_{-\ell}$ with z_ℓ^T , and we form the difference, so that the dominant term on the left-hand side becomes $-c_{\ell,1}ih^{-1}\frac{d}{dt}\|z_\ell\|^2$ (note $c_{-\ell,1} = -c_{\ell,1}$). Dividing by $-c_{\ell,1}ih^{-1}$ gives

$$\begin{aligned} \frac{i}{c_{\ell,1}h} \left(z_{-\ell}^T \frac{\rho}{\sigma}(\zeta_\ell e^{hD}) z_\ell - z_\ell^T \frac{\rho}{\sigma}(\zeta_{-\ell} e^{hD}) z_{-\ell} \right) \\ = \frac{ih}{c_{\ell,1}} \left(-z_{-\ell}^T \nabla_{z_{-\ell}} \mathcal{U}(\mathbf{z}) + z_\ell^T \nabla_{z_\ell} \mathcal{U}(\mathbf{z}) \right). \end{aligned} \quad (5.21)$$

We first estimate the right-hand expression. Since

$$\nabla_{z_{-\ell}} \mathcal{U}(\mathbf{z}) = \nabla^2 U(z_0) z_\ell + \mathcal{O}(\delta^2),$$

as long as (5.11) is satisfied, we obtain from the symmetry of the Hessian that the right-hand side of (5.21) is of size $\mathcal{O}(h\delta^3)$. The dominant $\mathcal{O}(h\delta^3)$ term is present only if $\zeta_{-\ell}$ can be written as the product of two roots of $\rho(\zeta)$ other than 1. If this is not the case, the expression (5.21) is of size $\mathcal{O}(h\delta^4)$.

Using the expansion (5.15) on the left-hand side of (5.21) and the relations (for $z = z_\ell$)

$$\begin{aligned} \operatorname{Re} \bar{z}^T z^{(2m+1)} &= \operatorname{Re} \frac{d}{dt} \left(\bar{z}^T z^{(2m)} - \dot{\bar{z}}^T z^{(2m-1)} \dots \mp \frac{1}{2} (\bar{z}^{(m)})^T z^{(m)} \right) \\ \operatorname{Im} \bar{z}^T z^{(2m+2)} &= \operatorname{Im} \frac{d}{dt} \left(\bar{z}^T z^{(2m+1)} - \dot{\bar{z}}^T z^{(2m)} + \dots \pm (\bar{z}^{(m)})^T z^{(m+1)} \right) \end{aligned}$$

we obtain that (5.21) is, up to $\mathcal{O}(h^N)$, the total derivative of a function depending on \mathbf{z} and its derivatives.

By construction the dominant term is $\frac{d}{dt}\|z_\ell\|^2$. The following terms have at least one more power of h and at least one derivative which by (3.25) gives rise to an additional factor h . Eliminating higher derivatives with the help of (3.25), we arrive at a function of the form

$$\mathcal{K}_\ell(y, \dot{y}, \mathbf{z}^*) = \|z_\ell\|^2 + h^2 K_{\ell,2}(y, \dot{y}, \mathbf{z}^*) + \dots + h^{N-1} K_{\ell,N-1}(y, \dot{y}, \mathbf{z}^*). \quad (5.22)$$

As we have seen, its total derivative is of size $\mathcal{O}(h\delta^3)$ or smaller. We summarize these considerations in the following theorem.

Theorem 5.4. *Along every solution of the truncated modified equation (3.25) the function $\mathcal{K}_\ell(y, \dot{y}, \mathbf{z}^*)$ satisfies for $1 < \ell < k$*

$$\mathcal{K}_\ell(y(t), \dot{y}(t), \mathbf{z}^*(t)) = \mathcal{K}_\ell(y(0), \dot{y}(0), \mathbf{z}^*(0)) + \mathcal{O}(th^N) + \mathcal{O}(th\delta^3) \quad (5.23)$$

as long as the solution stays in the set defined by (5.11). The second error term is replaced by $\mathcal{O}(th\delta^4)$ if no root of $\rho(\zeta)$ other than 1 is the product of two other roots. Moreover,

$$\mathcal{K}_\ell(y, \dot{y}, \mathbf{z}^*) = \|z_\ell\|^2 + \mathcal{O}(h^2\delta^2). \quad (5.24)$$

This result allows us to write the numerical solution in a form that is suitable for deriving long-time error estimates. Let us first collect the necessary assumptions:

- (A1) the multistep method (5.7) is symmetric, s -stable, and of order r ;
- (A2) the potential function $U(q)$ of (5.8) is defined and analytic in an open neighbourhood of a compact set K ;
- (A3) the starting approximations q_0, \dots, q_{k-1} are such that the initial values for (3.25) obtained from Lemma 3.7 satisfy $y(0) \in K$, $\|\dot{y}(0)\| \leq M$, and $\|z_\ell(0)\| \leq \delta/2$ for $1 < \ell < k$;
- (A4) the numerical solution $\{q_n\}$ stays for $0 \leq nh \leq T$ in a compact set K_0 which has a positive distance to the boundary of K .

Theorem 5.5 (Hairer & Lubich 2004). *Assume (A1)–(A4). For sufficiently small h and δ and for a fixed truncation index N (large enough such that $h^N = \mathcal{O}(\delta^4)$), there exist functions $y(t)$ and $z_\ell(t)$ on an interval of length*

$$T = \mathcal{O}((h\delta)^{-1})$$

such that

- $q_n = y(nh) + \sum_{\ell \in \mathcal{I}^*} \zeta_\ell^n z_\ell(nh)$ for $0 \leq nh \leq T$;
- on every subinterval $[jh, (j+1)h)$ the functions $y(t), z_\ell(t)$ are a solution of the system (3.25);
- the functions $y(t), z_\ell(t)$ have jump discontinuities of size $\mathcal{O}(h^{N+2})$ at the grid points jh ;
- $\|z_\ell(t)\| \leq \delta$ for $0 \leq t \leq T$.

If no root of $\rho(\zeta)$ other than 1 is the product of two other roots, all these estimates are valid on an interval of length $T = \mathcal{O}((h\delta^2)^{-1})$.

Proof. To define the functions $y(t), z_\ell(t)$ on the interval $[jh, (j+1)h)$ we consider the k consecutive numerical solution values $q_j, q_{j+1}, \dots, q_{j+k-1}$. We compute initial values for (3.25) according to Lemma 3.7, and we let $y(t), z_\ell(t)$ be a solution of (3.25) on $[jh, (j+1)h)$. Because their defect is $\mathcal{O}(h^N)$ and $\mathcal{O}(h^{N+1})$, respectively, such a construction yields jump discontinuities of size $\mathcal{O}(h^{N+2})$ at the grid points.

It follows from Theorem 5.4 that $\mathcal{K}_\ell(y(t), \dot{y}(t), \mathbf{z}^*(t))$ remains constant up to an error of size $\mathcal{O}(h^2\delta^3)$ on the interval $[jh, (j+1)h)$. Taking into account the jump discontinuities, we find that

$$\mathcal{K}_\ell(y(t), \dot{y}(t), \mathbf{z}^*(t)) \leq \mathcal{K}_\ell(y(0), \dot{y}(0), \mathbf{z}^*(0)) + C_1 th\delta^3 + C_2 th^{N+1} \quad (5.25)$$

as long as $\|z_\ell(t)\| \leq \delta$. By (5.24) this then implies

$$\|z_\ell(t)\|^2 \leq \|z_\ell(0)\|^2 + C_1 t h \delta^3 + C_2 t h^{N+1} + C_3 h^2 \delta^2. \quad (5.26)$$

The assumption $\|z_\ell(t)\| \leq \delta$ is certainly satisfied as long as $C_1 t h \delta \leq 1/4$, $C_2 t h^{N+1} \leq \delta^2/4$, and $C_3 h^2 \leq 1/4$, so that the right-hand side of (5.26) is bounded by δ^2 . This proves not only the estimate for $\|z_\ell(t)\|$, but at the same time it guarantees recursively that the above construction of the functions $y(t), z_\ell(t)$ is feasible. \square

Notice that for initial values computed by a sufficiently accurate one-step method the constant δ can be chosen as small as $\mathcal{O}(h^{r+2})$ where r is the order of the multistep method (cf. Lemma 3.7). The above estimates are therefore valid on very long time intervals.

Example 5.6. To illustrate the long-time behaviour of the parasitic terms z_ℓ we consider the pendulum equation $\ddot{q} = -\sin q$, and we apply the symmetric multistep methods (B) and (C) of Example 1.2. For method (C), the starting values are chosen far from a smooth solution, so that the propagation of the parasitic terms in the numerical solution can be better observed. We compute the velocity approximation by

$$v_n = \frac{h}{12} (8(q_{n+1} - q_{n-1}) - (q_{n+2} - q_{n-2})). \quad (5.27)$$

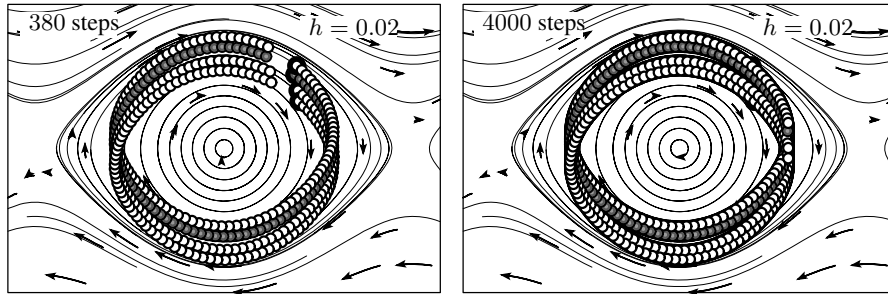


Fig. 5.2. Stable propagation of perturbations in the starting values for method (C) of Example 1.2; initial values are $q_0 = 1.141$, $q_1 = 1.158$, $q_2 = 1.178$, and $q_3 = 1.206$

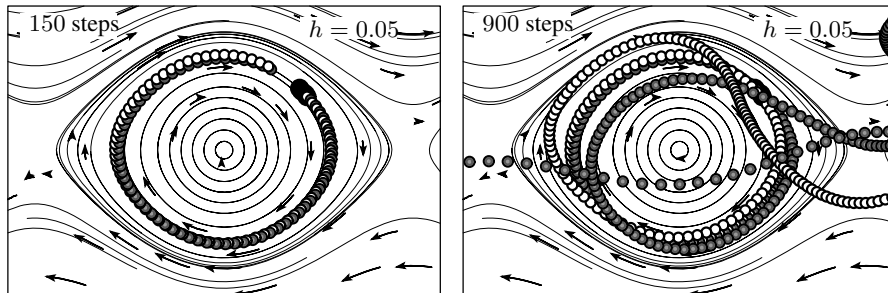


Fig. 5.3. Unstable propagation of perturbations in the starting values, for method (B) of Example 1.2; initial values are $q_0 = 1.147$, $q_1 = 1.183$, $q_2 = 1.255$, and $q_3 = 1.286$

Figure 5.2 shows the numerical solution (q_n, v_n) for $n \geq 2$. The values for $n = 2, 3, 4, 5$ are indicated by larger black bullets. The parasitic roots of method (C) are $\pm i$ and both are simple. The numerical solution is therefore of the form

$$q_n = y(nh) + i^n z_1(nh) + (-i)^n \overline{z_1(nh)} + (-1)^n z_2(nh).$$

One observes in Fig. 5.2 that the functions $z_j(t)$ not only remain bounded and small, but they stay essentially constant over the considered interval. This should be compared to Fig. 3.1, where the parasitic functions $z_j(t)$ are bounded, but not constant.

Method (B) has a double parasitic root at -1 and, therefore, is not s -stable. Its numerical solution behaves like $q_n = y(nh) + (-1)^n z(nh)$. In Fig. 5.3 every second approximation is drawn in grey. One sees that the numerical solution stays on two smooth curves $y(t) + z(t)$ and $y(t) - z(t)$ which, however, do not remain close to each other.

XV.6 Explanation of the Long-Time Behaviour

The bounds on the parasitic solution components of Sect. XV.5.3 allow us to get rigorous statements on the long-time behaviour of multistep methods (5.7) for second order differential equations. The following results are taken from Hairer & Lubich (2004). We do not know of similar results for multistep methods (1.1).

XV.6.1 Conservation of Energy and Angular Momentum

The energy conservation is now a direct consequence of Theorems 5.3 and 5.5. We shall use the representation of q_n in terms of functions $y(t), z_\ell(t)$ as in Theorem 5.5. Taking into account the jump discontinuities of these functions, Theorem 5.3 yields

$$\mathcal{H}(y(t), \dot{y}(t), \mathbf{z}^*(t)) = \mathcal{H}(y(0), \dot{y}(0), \mathbf{z}^*(0)) + \mathcal{O}(th^3\delta^4) + \mathcal{O}(th^{N+1}).$$

We have $\delta = \mathcal{O}(h^{r+1})$ if the starting approximations are computed by a r th order one-step method. If N is chosen sufficiently large, this together with (5.19) implies

$$H(y(t), \dot{y}(t)) = H(y(0), \dot{y}(0)) + \mathcal{O}(h^p) \quad \text{for } 0 \leq t \leq T = \mathcal{O}(h^{-p-2}).$$

If the velocity approximation $p_n = v_n$ is given by a r th order finite difference formula (3.11), it follows from Theorem 5.5 that $p_n = \dot{y}(nh) + \mathcal{O}(h^r)$ provided the truncation index N is sufficiently large. This proves the following result, and explains the excellent long-time behaviour of method (C) in Fig. 1.2.

Theorem 6.1 (Total Energy). *For a problem $\ddot{q} = -\nabla U(q)$ with total energy $H(p, q) = \frac{1}{2}p^T p + U(q)$, the numerical solution of an s -stable symmetric multistep method (5.7) of order r satisfies*

$$H(q_n, p_n) = H(q_0, p_0) + \mathcal{O}(h^r) \quad \text{for } nh \leq h^{-r-2}.$$

If no root of $\rho(\zeta)$ other than 1 is a product of two other roots, the statement holds on intervals of length $\mathcal{O}(h^{-2r-3})$. \square

We assume next that the differential equation $\ddot{q} = -\nabla U(q)$ has a quadratic first integral of the form $L(q, \dot{q}) = \dot{q}^T A q$ (e.g., the angular momentum in N -body problems). This means that A is skew-symmetric and $\nabla U(q)^T A q = 0$. The last equation can also be interpreted as the invariance relation $U(e^{\tau A} q) = U(q)$. This property implies for $\mathcal{U}(\mathbf{z})$, given by (5.9), that $\mathcal{U}(e^{\tau A} \mathbf{z}) = \mathcal{U}(\mathbf{z})$ (here $e^{\tau A} \mathbf{z} = (e^{\tau A} z_\ell)_{\ell \in \mathcal{I}}$). Along solutions $\mathbf{z}(t)$ of the modified equations (5.10) we therefore have up to terms of size $\mathcal{O}(h^N)$

$$0 = \left. \frac{d}{d\tau} \right|_{\tau=0} \mathcal{U}(e^{\tau A} \mathbf{z}) = \sum_{\ell \in \mathcal{I}} z_{-\ell}^T A \nabla_{z_{-\ell}} \mathcal{U}(\mathbf{z}) = \sum_{\ell \in \mathcal{I}} h^{-2} z_{-\ell}^T A \left(\frac{\rho}{\sigma} \right) (\zeta_\ell e^{hD}) z_\ell.$$

If $\sigma(\zeta)$ has a root ζ_ℓ , then the corresponding term is omitted from the last sum, leading to a remainder term which in the worst case is $\mathcal{O}(h^3 \delta^4)$, as in Theorem 5.3. Like in the previous proofs, the last sum is, for skew-symmetric A , the total derivative of a function

$$\mathcal{L}(y, \dot{y}, \mathbf{z}^*) = L_0(y, \dot{y}, \mathbf{z}^*) + \dots + h^{N-1} L_{N-1}(y, \dot{y}, \mathbf{z}^*)$$

which satisfies (under the same assumptions as in Theorem 5.3)

$$\mathcal{L}(y(t), \dot{y}(t), \mathbf{z}^*(t)) = \mathcal{L}(y(0), \dot{y}(0), \mathbf{z}^*(0)) + \mathcal{O}(th^3 \delta^4) + \mathcal{O}(th^{N+1})$$

and

$$\mathcal{L}(y, \dot{y}, \mathbf{z}^*) = L(y, \dot{y}) + \mathcal{O}(h^p) + \mathcal{O}(\delta^2/h). \quad (6.1)$$

We therefore obtain the following result.

Theorem 6.2 (Angular Momentum). *Let $L(q, \dot{q}) = \dot{q}^T A q$ be a first integral of $\ddot{q} = -\nabla U(q)$. The numerical solution of an s -stable symmetric multistep method (5.7) of order r then satisfies*

$$L(q_n, p_n) = L(q_0, p_0) + \mathcal{O}(h^r) \quad \text{for } nh \leq h^{-r-2}.$$

If no root of $\rho(\zeta)$ other than 1 is a product of two other roots, the statement holds on intervals of length $\mathcal{O}(h^{-2r-3})$. \square

XV.6.2 Linear Error Growth for Integrable Systems

The differential equation $\ddot{q} = -\nabla U(q)$, written as $\dot{q} = v, \dot{v} = -\nabla U(q)$, is reversible with respect to the involution $v \mapsto -v$. Assume that it is also an integrable system in the sense of Definition XI.1.1, and denote by $a = I(q, v)$ the action variables, and by $\omega(a)$ the frequencies of the system.

By Theorem 5.5, the numerical solution can be written as $q_n = y(nh) + \sum_{\ell \in \mathcal{I}^*} \zeta_\ell^n z_\ell(nh)$, where (at least locally) $y(t)$ is the solution of a modified differential equation (first equation of (3.25))

$$\ddot{y} = f_{0,0}(y, \dot{y}, \mathbf{z}^*) + h f_{0,1}(y, \dot{y}, \mathbf{z}^*) + \dots + h^{N-1} f_{0,N-1}(y, \dot{y}, \mathbf{z}^*) \quad (6.2)$$

which, for $\mathbf{z}^* = 0$ becomes the reversible modified differential equation (3.9). Since $z_j(t) = \mathcal{O}(\delta)$ (see Theorem 5.5) and since \mathbf{z}^* appears at least quadratically in (6.2), this equation is a $\mathcal{O}(\delta^2)$ perturbation of (3.9). We are now in the position to apply the results of Lemma XI.2.1 and Theorem XI.3.1. The additional (non-reversible) perturbation of size $\mathcal{O}(\delta^2)$ in the differential equation (6.2) produces an error term of size $\mathcal{O}(t\delta^2)$ in the action variables and of size $\mathcal{O}(t^2\delta^2)$ in the angle variables. If $\delta = \mathcal{O}(h^{r+1})$, these terms are negligible with respect to those already appearing in Theorem XI.3.1. The errors due to the jump discontinuities (Theorem 5.5) are also negligible. We have thus proved the following statement.

Theorem 6.3. *Consider applying the s -stable symmetric multistep method (5.7) of order r to an integrable reversible system $\ddot{q} = -\nabla U(q)$ with real-analytic potential U . Suppose that $\omega^* \in \mathbb{R}^d$ satisfies the diophantine condition (X.2.4). Then, there exist positive constants C, c and h_0 such that the following holds for all step sizes $h \leq h_0$: every numerical solution (q_n, v_n) starting with frequencies $\omega_0 = \omega(I(q_0, v_0))$ such that $\|\omega_0 - \omega^*\| \leq c|\log h|^{-\nu-1}$, satisfies*

$$\begin{aligned} \|(q_n, v_n) - (q(t), v(t))\| &\leq C t h^r \\ \|I(q_n, v_n) - I(q_0, v_0)\| &\leq C h^r \end{aligned} \quad \text{for } 0 \leq t = nh \leq h^{-r}.$$

The constants h_0, c, C depend on d, γ, ν and on bounds of the potential. \square

XV.7 Practical Considerations

In computations with multistep methods one can observe resonance phenomena, if relatively large step sizes are used. This and the use of variable step sizes are the subject of this section.

XV.7.1 Numerical Instabilities and Resonances

Soon after Quinlan and Tremaine's methods were published, however, Alar Toomre discovered a disturbing feature of the methods, . . .
(G.D. Quinlan 1999)

It is a simple task to derive multistep methods of high order. Consider, for example, methods of the form (1.8) for second order differential equations $\ddot{y} = f(y)$. Their order is determined by the condition (1.9). We choose arbitrarily $\rho(\zeta)$ such that $\zeta = 1$ is a double zero and the stability condition is satisfied. Condition (1.9) then gives

$$\sigma(\zeta) = \rho(\zeta)/\log^2 \zeta + \mathcal{O}((\zeta - 1)^r).$$

Expanding the right-hand expression into a Taylor series at $\zeta = 1$ and truncating suitably, this yields the corresponding σ polynomial. If we take

$$\rho(\zeta) = (\zeta - 1)^2(\zeta^6 + \zeta^4 + \zeta^3 + \zeta^2 + 1), \quad (7.1)$$

Table 7.1. Symmetric multistep methods for second order problems; $k = 8$ and order $r = 8$

i	SY8		SY8B		SY8C	
	α_i	$12096 \beta_i$	α_i	$120960 \beta_i$	α_i	$8640 \beta_i$
0	1	0	1	0	1	0
1	-2	17671	0	192481	-1	13207
2	2	-23622	0	6582	0	-8934
3	-1	61449	$-1/2$	816783	0	42873
4	0	-50516	-1	-156812	0	-33812

we get in this way Method SY8 of Table 7.1, a method proposed by Quinlan & Tremaine (1990) for computations in celestial mechanics. All methods of Table 7.1 are 8-step methods, of order 8, and symmetric, i.e., the relations $\alpha_i = \alpha_{k-i}$ and $\beta_i = \beta_{k-i}$ are satisfied. Therefore, we present the coefficients only for $i \leq k/2$.

These methods give approximations y_n to the solution of the differential equation. If also derivative approximations are needed, we get them by finite differences, e.g., for the 8th order methods of Table 7.1 we use

$$\dot{y}_n = \frac{1}{840h} \left(672 (y_{n+1} - y_{n-1}) - 168 (y_{n+2} - y_{n-2}) + 32 (y_{n+3} - y_{n-3}) - 3 (y_{n+4} - y_{n-4}) \right). \quad (7.2)$$

We apply this method to the Kepler problem (I.2.2), once with eccentricity $e = 0$ and once with $e = 0.2$, and initial values (I.2.11), such that the period of the exact solution is 2π . Starting approximations are computed accurately with a high order Runge–Kutta method. We apply Method SY8 with many different step sizes ranging from $2\pi/30$ to $2\pi/95$, and we plot in Fig. 7.1 the maximum error of the total energy as a function of $2\pi/h$ (where h denotes the step size). We see that in general the error decreases with the step size, but there is an extremely large error for $h \approx 2\pi/60$. For $e \neq 0$, further peaks can be observed at integral multiples of 5 and 6. It is our aim to understand this behaviour.

Instabilities. We put $z = q_1 + iq_2$, so that the Kepler problem becomes

$$\ddot{z} = \psi(|z|)z, \quad \psi(r) = -r^{-3},$$

and we choose initial values such that $z(t) = e^{it}$ is a circular motion (eccentricity $e = 0$). The numerical solution of (1.8) is therefore defined by the relation

$$\sum_{j=0}^k \alpha_j z_{n+j} = h^2 \sum_{j=0}^k \beta_j \psi(|z_{n+j}|) z_{n+j}. \quad (7.3)$$

Approximating $\psi(|z_{n+j}|)$ with $\psi(1) = -\omega^2$, we get a linear recurrence relation with characteristic polynomial

$$S(\omega h, \zeta) = \rho(\zeta) + \omega^2 h^2 \sigma(\zeta).$$

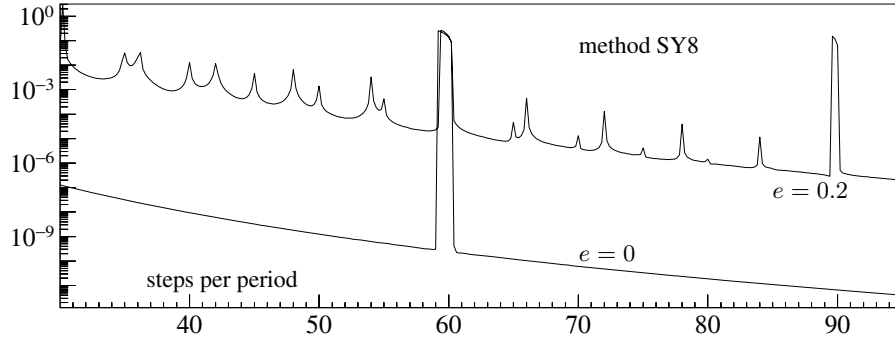


Fig. 7.1. Maximum error in the total energy during the integration of 2500 orbits of the Kepler problem as a function of the number of steps per period

The principal roots of $S(\omega h, \zeta) = 0$ satisfy $\zeta_1(\omega h) \approx e^{i\omega h}$ and $\zeta_2(\omega h) \approx e^{-i\omega h}$, and we have $|\zeta_j(\omega h)| = 1$ for all j and for sufficiently small h , because the method is symmetric (Exercise 2). As a consequence of $|\zeta_1(\omega h)| = 1$, the values $\hat{z}_n := \zeta_1(\omega h)^n$ are not only a solution of the linear recurrence relation, but also of the nonlinear relation (7.3). Our aim is to study the stability of this numerical solution. We therefore consider a perturbed solution

$$z_n = \zeta_1(\omega h)^n (1 + u_n).$$

Using $|z_n| = 1 + \frac{1}{2}(u_n + \bar{u}_n) + \mathcal{O}(|u_n|^2)$ and neglecting the quadratic and higher order terms of $|u_n|$ in the relation (7.3), we get

$$\sum_{j=0}^k (\alpha_j + \omega^2 h^2 \beta_j) \zeta_1(\omega h)^j u_{n+j} = \frac{h^2}{2} \psi'(1) \sum_{j=0}^k \beta_j \zeta_1(\omega h)^j (u_{n+j} + \bar{u}_{n+j}).$$

Considering also the complex conjugate of this relation, and eliminating \bar{u}_{n+j} , we obtain a linear recurrence relation for u_n with characteristic polynomial

$$S(\omega h, \zeta_1(\omega h)\zeta) \cdot S(\omega h, \zeta_1(\omega h)^{-1}\zeta) + \mathcal{O}(h^2). \quad (7.4)$$

For small h , its zeros are close to $\zeta_1(\omega h)^{-1}\zeta_j$ and $\zeta_1(\omega h)\zeta_l$. If two of these zeros collapse, the $\mathcal{O}(h^2)$ terms in (7.4) can produce a root of modulus larger than one, so that instability occurs. This is the case, if two roots ζ_j, ζ_l of $\rho(\zeta) = 0$ satisfy $\zeta_j \zeta_l^{-1} \approx \zeta_1^2 \approx e^{2i\omega h}$, or

$$\theta_j - \theta_l = \frac{4\pi}{N}, \quad (7.5)$$

where $\zeta_j = e^{i\theta_j}$ and $h = 2\pi/N$.

For the Method SY8 of Table 7.1, the spurious zeros of $\rho(\zeta)$ have arguments $\pm 4\pi/5, \pm 2\pi/5$, and $\pm 2\pi/6$. With $\theta_j = 2\pi/5$ and $\theta_l = 2\pi/6$, the condition (7.5) gives $N = 60$ as a candidate for instability. This explains the experiment of Fig. 7.1 for $e = 0$. A study of the stability of orbits with eccentricity $e \neq 0$ (see Quinlan

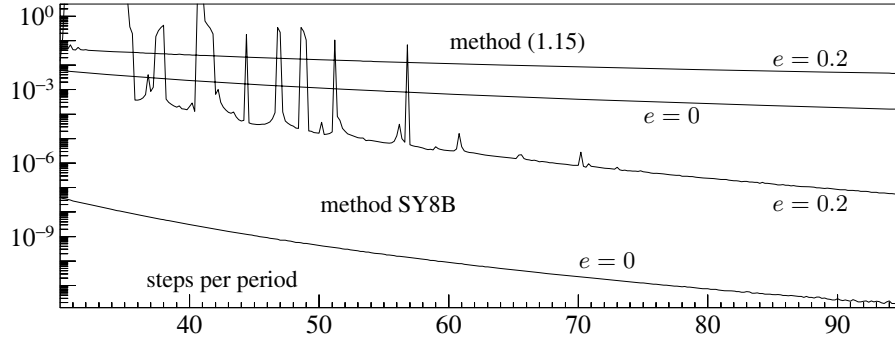


Fig. 7.2. Maximum error in the total energy during the integration of 2500 orbits of the Kepler problem as a function of the number of steps per period

1999) shows that instabilities can also occur when $4\pi/N$ is replaced with $2q\pi/N$ ($q = 2, 3, \dots$) in the relation (7.5).

To avoid these instabilities as far as possible, Quinlan (1999) constructed symmetric multistep methods, where the spurious roots of $\rho(\zeta) = 0$ are well spread out on the unit circle and far from $\zeta = 1$. As a result he proposes Method SY8B of Table 7.1. The same experiment as above yields the results of Fig. 7.2. The ρ -polynomial of Method SY8B is

$$\rho(\zeta) = (\zeta - 1)^2(\zeta^6 + 2\zeta^5 + 3\zeta^4 + 3.5\zeta^3 + 3\zeta^2 + 2\zeta + 1),$$

and the θ_j of the spurious roots are $\pm 2\pi/2.278$, $\pm 2\pi/3.353$, and $\pm 2\pi/4.678$. The condition (7.5) is satisfied only for $N \leq 23.67$, which implies that no instability occurs for $e = 0$ in the region of the experiment of Fig. 7.2.

To illustrate the importance of high order methods, we included in Fig. 7.2 the results of the second order partitioned multistep method (1.15).

XV.7.2 Extension to Variable Step Sizes

Variable step size multistep methods for second order differential equations $\ddot{y} = f(y)$ are of the form

$$\sum_{j=0}^k \alpha_j(h_n, \dots, h_{n+k-1}) y_{n+j} = h_{n+k-1}^2 \sum_{j=0}^k \beta_j(h_n, \dots, h_{n+k-1}) f(y_{n+j}),$$

where the coefficients α_j and β_j are allowed to depend on the step sizes h_n, \dots, h_{n+k-1} , more precisely, on the ratios $h_{n+1}/h_n, \dots, h_{n+k-1}/h_{n+k-2}$. They yield approximations y_n to $y(t_n)$ on a variable grid given by $t_{n+1} = t_n + h_n$. Such a method is of order r (cf. formula (1.9)), if

$$\sum_{j=0}^k \alpha_j(h_n, \dots, h_{n+k-1}) y(t_{n+j}) = h_{n+k-1}^2 \sum_{j=0}^k \beta_j(h_n, \dots, h_{n+k-1}) \ddot{y}(t_{n+j}) \quad (7.6)$$

for all polynomials $y(t)$ of degree $\leq r + 1$. It is *stable*, if the ρ -polynomial with coefficients $\alpha_j(h, \dots, h)$ (constant step size) satisfies the stability condition of Sect. XV.1.2 (see Theorem III.5.7 of Hairer, Nørsett & Wanner (1993) and Cano & Durán (2003a)).

All methods of Sect. XV.7.1 can be extended to symmetric, variable step size integrators. This has been discovered by Cano & Durán (2003b). For clarity of notation we let $\tilde{\alpha}_j, \tilde{\beta}_j$ ($j = 0, \dots, k$) be the coefficients of such a fixed step size method. Cano & Durán propose putting

$$\beta_j(h_n, \dots, h_{n+k-1}) = \frac{h_n}{h_{n+k-1}} \tilde{\beta}_j, \quad (7.7)$$

and to determine $\alpha_j(h_n, \dots, h_{n+k-1})$ such that symmetry and order $k - 2$ (for arbitrary step sizes) are achieved. We also suppose (7.7), but we determine the coefficients $\alpha_j(h_n, \dots, h_{n+k-1})$ such that (7.6) holds for all polynomials $y(t)$ of degree $\leq k$. This uniquely determines these coefficients whenever $h_n > 0, \dots, h_{n+k-1} > 0$ (Vandermonde type system) and gives the following properties.

Lemma 7.1. *For even k , let $(\tilde{\alpha}_j, \tilde{\beta}_j)$ define a symmetric, stable k -step method (1.8) of order k , and consider the variable step size method given by (7.7) and $\alpha_j(h_n, \dots, h_{n+k-1})$ such that (7.6) holds for all polynomials y satisfying $\deg y \leq k$. This method extends the fixed step size formula, i.e.,*

$$\alpha_j(h, \dots, h) = \tilde{\alpha}_j, \quad \beta_j(h, \dots, h) = \tilde{\beta}_j, \quad (7.8)$$

it satisfies the symmetry relations

$$\begin{aligned} \alpha_j(h_n, \dots, h_{n+k-1}) &= \alpha_{k-j}(h_{n+k-1}, \dots, h_n) \\ h_{n+k-1}^2 \beta_j(h_n, \dots, h_{n+k-1}) &= h_n^2 \beta_{k-j}(h_{n+k-1}, \dots, h_n), \end{aligned} \quad (7.9)$$

and it is of order $k - 1$ for arbitrary step sizes. Moreover, it behaves like a method of order k , if $h_{n+1} = h_n(1 + \mathcal{O}(h_n))$ uniformly in n .

Proof. The relation (7.8) for β_j follows at once from (7.7), and for α_j it is a consequence of the uniqueness of the solution of the linear system for the α_j .

The second condition of (7.9) follows directly from (7.7) and from the symmetry of the underlying fixed step size method ($\tilde{\beta}_{k-j} = \tilde{\beta}_j$ for all j). Inserting (7.7) into (7.6), replacing $y(t)$ with $y(t_{n+k} + t_n - t)$, and reversing the order of h_n, \dots, h_{n+k-1} yields

$$\sum_{j=0}^k \alpha_j(h_{n+k-1}, \dots, h_n) y(t_{n+k-j}) = h_n h_{n+k-1} \sum_{j=0}^k \tilde{\beta}_j \ddot{y}(t_{n+k-j}).$$

Using $\tilde{\beta}_{k-j} = \tilde{\beta}_j$ this shows that $\alpha_{k-j}(h_{n+k-1}, \dots, h_n)$ satisfies exactly the same linear system as $\alpha_j(h_n, \dots, h_{n+k-1})$, so that also the first relation of (7.9) is verified.

By definition, the variable step size method is at least of order $k - 1$. Under the assumption $h_{n+1} = h_n(1 + \mathcal{O}(h_n))$ the defect in (7.6) is of the form

$$h_n^{k+1}D(h_n, \dots, h_{n+k-1}) = h_n^{k+1}D(h_n, \dots, h_n) + \mathcal{O}(h_n^{k+2})$$

for all sufficiently smooth $y(t)$. Since the constant step size method is of order k , the expression $D(h_n, \dots, h_n)$ is of size $\mathcal{O}(h_n)$, so that we observe convergence of order k . \square

The symmetry relation (7.9) has the following interpretation: if the approximations y_n, \dots, y_{n+k-1} used with step sizes h_n, \dots, h_{n+k-1} yield y_{n+k} , then the values y_{n+k}, \dots, y_{n+1} applied with h_{n+k-1}, \dots, h_n yield y_n as a result (since the coefficients α_j and β_j only depend on step size ratios and the multistep formula only on h_{n+k-1}^2 , the same result is obtained with $-h_{n+k-1}, \dots, -h_n$). This is the analogue of the definition of symmetry for one-step methods.

For obtaining a good long-time behaviour, the step sizes also have to be chosen in a symmetric and reversible way (see Sect. VIII.3). One possibility is to take step sizes

$$h_{n+k-1} = \frac{\varepsilon}{2} \left(\sigma(y_{n+k-1}) + \sigma(y_{n+k}) \right), \quad (7.10)$$

where $\varepsilon > 0$, and $\sigma(y)$ is a given positive monitor function. This condition is an implicit equation for h_{n+k-1} , because y_{n+k} depends on h_{n+k-1} . It has to be solved iteratively. Notice, however, that for an explicit multistep formula no further force evaluations are necessary during this iteration. Such a choice of the step size guarantees that whenever h_{n+k-1} is chosen when stepping from y_n, \dots, y_{n+k-1} with h_n, \dots, h_{n+k-2} to y_{n+k} , the step size h_n is chosen when stepping backwards from y_{n+k}, \dots, y_{n+1} with $h_{n+k-1}, \dots, h_{n+1}$ to y_n .

Implementation. For given initial values y_0, \dot{y}_0 , the starting approximations y_1, \dots, y_{k-1} should be computed accurately (for example, by a high-order Runge–Kutta method) with step sizes satisfying (7.10). The solution of the scalar nonlinear equation (7.10) has to be done carefully in order to reduce the overhead of the method. In our code we use $h_{n+k-1} := h_{n+k-2}^2 / h_{n+k-3}$ as predictor, and we apply modified Newton iterations with the derivative approximated by finite differences.

The coefficients $\alpha_j(h_n, \dots, h_{n+k-1})$ have to be computed anew in every iteration. We use the basis

$$p_i(t) = \prod_{j=0}^{i-1} (t - t_{n+j}), \quad i = 0, \dots, k$$

for the polynomials of degree $\leq k$ in (7.6). This leads to a linear triangular system for $\alpha_0, \dots, \alpha_k$. As noticed by Cano & Durán (2003b), the coefficients $p_i(t_j)$ and $\ddot{p}_i(t_j)$ can be obtained efficiently from the recurrence relations

$$\begin{aligned} p_0(t) &= 1, & p_{i+1}(t) &= (t - t_i)p_i(t) \\ \dot{p}_0(t) &= 0, & \dot{p}_{i+1}(t) &= (t - t_i)\dot{p}_i(t) + p_i(t) \\ \ddot{p}_0(t) &= 0, & \ddot{p}_{i+1}(t) &= (t - t_i)\ddot{p}_i(t) + 2\dot{p}_i(t). \end{aligned}$$

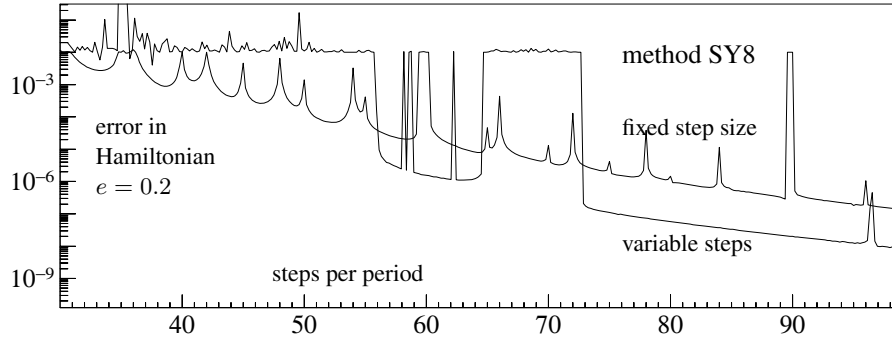


Fig. 7.3. Maximum error in the total energy during the integration of 2500 orbits of the Kepler problem as a function of the number of steps per period

During the iterations for the solution of the nonlinear equation (7.10) only the values of $p_i(t_{n+k})$ have to be updated.

Numerical Experiment. We repeat the experiment of Fig. 7.1 with the method SY8, but this time in the variable step size version and with $\sigma(y) = \|y\|^2$ as step size monitor. We have computed 2500 periods of the Kepler problem with eccentricity $e = 0.2$, and we have plotted in Fig. 7.3 the maximal error in the Hamiltonian as a function of the number of steps per period (for a comparison we have also included the result of the fixed step size implementation). Similar to (7.2) we use approximations \dot{y}_n that are the derivative of the interpolation polynomial passing through $y_n, y_{n+1}, y_{n+2}, \dots$ such that the correct order is obtained. The computation is stopped when the error exceeds 10^{-2} .

As expected, the error is smaller for the variable step size version, and it is seen that the peaks due to numerical resonances are now much less although they are not completely removed. For large step sizes, the performance deteriorates, but this is not a serious problem, because these methods are recommended only for high accuracy computations.

It should be remarked that the overhead, due to the computation of the coefficients α_j and the solution of the nonlinear equation (7.10), is rather high. Therefore, the use of variable step sizes is recommended only when force evaluations $f(y)$ are expensive or when constant step sizes are not appropriate. Cano & Durán (2003b) report an excellent performance of symmetric, variable step size multistep methods for computations of the outer solar system.

Despite the resonances and instabilities, then, symmetric methods can still be a better choice than Störmer methods for long integrations of planetary orbits provided that the user is aware of the dangers.

(G.D. Quinlan 1999)

XV.8 Multi-Value or General Linear Methods

General linear methods is a class of integration methods that covers Runge–Kutta as well as multistep methods. It is therefore of interest to study which of the results on the long-time behaviour can be extended.

So-called multi-value or general linear methods are defined by $Y_{n+1} = G_h(Y_n)$, where

$$\begin{aligned} Y_{n+1} &= DY_n + hBf(U_{n+1}) \\ U_{n+1} &= CY_n + hAf(U_{n+1}) \end{aligned} \quad (8.1)$$

with $f(U_{n+1}) = (f(u_{n+1}^1), \dots, f(u_{n+1}^s))^T$ for $U_{n+1} = (u_{n+1}^1, \dots, u_{n+1}^s)^T$, and $Y_n = (y_n^1, \dots, y_n^k)$. We use a sloppy notation in the sense that the matrices D, B, \dots should be replaced with $D \otimes I, B \otimes I, \dots$. For a computation, a starting procedure S_h and a finishing procedure F_h , which extracts the numerical approximation y_n from Y_n , have to be added (see Fig. 8.1). We assume the existence of a vector e such that with $\mathbb{1} = (1, \dots, 1)^T$

$$De = e, \quad Ce = \mathbb{1} \quad (8.2)$$

holds (preconsistency conditions). The vector Y_n is then an approximation to $ey(t_n)$ (more precisely to $e \otimes y(t_n)$).

For Runge–Kutta methods, $D = (1)$ is the one-dimensional identity, $B = (b_1, \dots, b_s)$, $C = \mathbb{1}$, and A is the usual Runge–Kutta matrix. For multistep methods, we have $Y_n = (y_{n+k-1}, \dots, y_n)^T$, and D is the $k \times k$ matrix with characteristic polynomial $\rho(\zeta)$ as in (2.1). For a detailed treatment of general linear methods we refer the reader to Chap. 4 of the monograph of Butcher (1987), and to Chap. III.8 of Hairer, Nørsett & Wanner (1993).

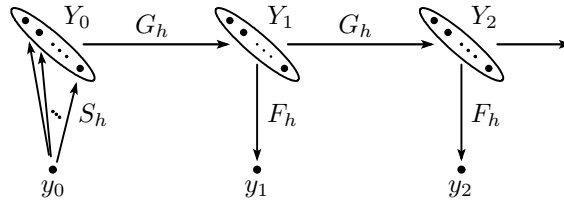


Fig. 8.1. Illustration of a multi-value method $Y_{n+1} = G_h(Y_n)$ with starting procedure S_h and finishing procedure F_h

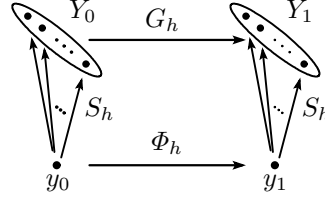
XV.8.1 Underlying One-Step Method and Backward Error Analysis

In analogy to multistep methods, a method (8.1) is called strictly stable, if all eigenvalues of D are inside the unit circle with the exception of the single eigenvalue $\zeta = 1$. An extension of Kirchgraber's result (Theorem 2.1) to strictly stable general linear methods is given by Stoffer (1993).

Theorem 8.1. Consider a strictly stable general linear method $Y_{n+1} = G_h(Y_n)$, and a finishing procedure $y_n = F_h(Y_n) = d^T Y_n + \mathcal{O}(h)$. Assume that (8.2) and $d^T e = 1$ hold.

(i) Then there exist a unique one-step method $\Phi_h(y)$ and a unique starting procedure $S_h(y)$ such that $G_h \circ S_h = S_h \circ \Phi_h$ and $F_h \circ S_h = Id$ hold.

(ii) The manifold $\mathcal{M}_h = \{S_h(y); y \in \mathbb{R}^d\}$ is invariant under G_h , and it is exponentially attractive.



Proof. Since the method is strictly stable, there exists a matrix T such that

$$T^{-1}DT = \begin{pmatrix} 1 & 0 \\ 0 & D_0 \end{pmatrix} \quad \text{with} \quad \|D_0\| < 1,$$

and $Te_1 = e$ (where $e_1 = (1, 0, \dots, 0)^T$). The proof closely follows that of Theorem 2.1. With the transformation $(\xi_n, \eta_n)^T = Z_n = T^{-1}Y_n$, the general linear method (8.1) becomes

$$\begin{pmatrix} \xi_{n+1} \\ \eta_{n+1} \end{pmatrix} = \begin{pmatrix} \xi_n \\ D_0 \eta_n \end{pmatrix} + hT^{-1}Bf(U_{n+1}). \quad (8.3)$$

with $U_{n+1} = CTZ_n + hAf(U_{n+1})$. As before, Theorem XII.3.1 can be applied and yields the existence of an attractive manifold $\mathcal{N}_h = \{(\xi, s(\xi)); \xi \in \mathbb{R}^d\}$, which is invariant under the mapping (8.3). We now invert the restriction of F_h onto the manifold \mathcal{N}_h . Due to $d^T e = 1$ and $Te_1 = e$, we have for $Z = Z(\xi) = (\xi, s(\xi))^T$ that

$$y = F_h(TZ(\xi)) = d^T TZ(\xi) + \dots = \xi + g(\xi), \quad (8.4)$$

where $g(\xi)$ is Lipschitz continuous with constant $\mathcal{O}(h)$. By the Banach fixed-point theorem the equation (8.4) has a unique solution $\xi = r(y)$. Putting

$$S_h(y) = TZ(r(y)) = T \begin{pmatrix} r(y) \\ s(r(y)) \end{pmatrix},$$

we have found the unique starting procedure satisfying $F_h \circ S_h = Id$ and $T^{-1}S_h(y) \in \mathcal{N}_h$. We finally define $\Phi_h = F_h \circ G_h \circ S_h$ and $\mathcal{M}_h = \{TZ; Z \in \mathcal{N}_h\}$, so that all statements of the theorem are verified. \square

It is our aim to extend the concept of an underlying one-step method to nearly all (including weakly stable) general linear methods.

Theorem 8.2. Consider a general linear method (8.1), and assume that $\zeta = 1$ is a single eigenvalue of the propagation matrix D . Furthermore, let $G_h(Y)$ and $F_h(Y) = d^T Y + \dots$ have expansions in powers of h , and assume that (8.2) and $d^T e = 1$ hold. Then there exist a unique formal one-step method

$$\Phi_h(y) = y + h d_1(y) + h^2 d_2(y) + \dots$$

and a unique formal starting procedure

$$S_h(y) = e y + h S_1(y) + h^2 S_2(y) + \dots,$$

such that formally $G_h \circ S_h = S_h \circ \Phi_h$ and $F_h \circ S_h = Id$ hold.

Proof. Expanding $S_h(\Phi_h(y))$ and $G_h(S_h(y))$ into powers of h , a comparison of the coefficients yields

$$e d_j(y) + (I - D) S_j(y) = \dots, \quad (8.5)$$

where a right-hand side depends on known functions and on $d_i(y), S_i(y)$ with $i < j$. Similarly, the condition $F_h(S_h(y)) = y$ leads to

$$d^T S_j(y) = \dots. \quad (8.6)$$

Due to the fact that $\zeta = 1$ is a single eigenvalue of D , and that $d^T e \neq 0$, the system (8.5)-(8.6) uniquely determines $d_j(y)$ and $S_j(y)$. \square

Backward Error Analysis for Smooth Numerical Solutions. The formal analysis of Chap. IX can be directly applied to the underlying one-step method of Theorem 8.2. This yields a modified differential equation, but only for the smooth numerical solution (cf. Sect. XV.3.1). Notice that this modified equation depends on the choice of the finishing procedure F_h .

XV.8.2 Symplecticity and Symmetry

Before giving a precise meaning to the symplecticity and symmetry of general linear methods, we establish the following lemma.

Lemma 8.3. For a general linear method $Y_{n+1} = G_h(Y_n)$ we consider two different finishing procedures $y_n = F_h(Y_n)$ and $\hat{y}_n = \hat{F}_h(Y_n)$:

$$\begin{array}{ccccccc} \hat{y}_0 & \xrightarrow{\hat{\Phi}_h} & \hat{y}_1 & \xrightarrow{\hat{\Phi}_h} & \hat{y}_2 & \xrightarrow{\hat{\Phi}_h} & \dots \\ \hat{S}_h \updownarrow \hat{F}_h & & \updownarrow \hat{F}_h & & \updownarrow \hat{F}_h & & \\ Y_0 & \xrightarrow{G_h} & Y_1 & \xrightarrow{G_h} & Y_2 & \xrightarrow{G_h} & \dots \\ S_h \updownarrow F_h & & \downarrow F_h & & \downarrow F_h & & \\ y_0 & \xrightarrow{\Phi_h} & y_1 & \xrightarrow{\Phi_h} & y_2 & \xrightarrow{\Phi_h} & \dots \end{array}$$

The two corresponding one-step methods $\Phi_h(y)$ and $\hat{\Phi}_h(y)$ (given by Theorem 8.2) are then conjugate to each other, i.e.,

$$\alpha_h^{-1} \circ \Phi_h \circ \alpha_h = \hat{\Phi}_h \quad \text{with} \quad \alpha_h = F_h \circ \hat{S}_h. \quad (8.7)$$

Proof. The equations involving the underlying one-step methods or the starting procedures have to be understood in the sense of formal series. By Theorem 8.2 we have $S_h(y) = ey + \mathcal{O}(h)$ and also $\widehat{S}_h(y) = ey + \mathcal{O}(h)$. It thus follows from $F_h \circ S_h = Id$ that $\alpha_h(y)$ is $\mathcal{O}(h)$ -close to the identity and therefore invertible. \square

The transformation α_h in the phase space is $\mathcal{O}(h)$ -close to the identity. The relation $\alpha_h^{-1} \circ \widehat{\Phi}_h^n \circ \alpha_h = \widehat{\Phi}_h^n$, which is a consequence of (8.7), therefore implies that the numerical solutions of Φ_h and $\widehat{\Phi}_h$ remain $\mathcal{O}(h)$ -close for all times. This means that the long-time behaviour of both methods is exactly the same.

Consequently, for a given general linear method G_h , it is sufficient to require symplecticity or symmetry for *one* finishing procedure only.

Definition 8.4 (Symplecticity). A general linear method G_h is called *symplectic* if there exists a finishing procedure F_h such that the underlying one-step method Φ_h of Theorem 8.2 is symplectic, i.e., $\Phi'_h(y)^T J \Phi'_h(y) = J$ in the sense of formal series.

The study of symplecticity of linear multistep methods (Sect. XV.4.1) was rather disappointing. We could not find one linear multistep method whose underlying one-step method is symplectic. For general linear methods, some necessary conditions for the symplecticity of the underlying one-step method are known which are hard to satisfy (Hairer & Leone 1998). For the moment, no symplectic general linear method (not equivalent to a one-step method) is known, and we conjecture that such a method does not exist, even in the class of partitioned general linear methods (treating the p and q variables by different methods).

After the disappointing non-existence conjecture of symplectic multi-value methods, we turn our attention to symmetric methods. We know from the previous chapters that for reversible Hamiltonian systems, the long time behaviour of symmetric one-step methods can be as good as that for symplectic methods. There are several definitions of symmetric general linear methods in the literature. However, they are either tailored to very special situations (e.g., Hairer, Nørsett & Wanner 1993), or they do not allow the proof of results that are expected to hold for symmetric methods.

Definition 8.5 (Symmetry). A general linear method G_h is called *symmetric* if there exists a finishing procedure F_h such that the underlying one-step method Φ_h of Theorem 8.2 is symmetric, i.e., $\Phi_{-h}(y) = \Phi_h^{-1}(y)$ in the sense of formal series.

Example 8.6. Consider the trapezoidal method in the role of G_h and the explicit Euler method with step size $-\gamma h$ as finishing procedure:

$$\begin{aligned} G_h : \quad Y_{n+1} &= Y_n + \frac{h}{2} \left(f(Y_n) + f(Y_{n+1}) \right) \\ F_h : \quad y_{n+1} &= Y_{n+1} - \gamma h f(Y_{n+1}) \end{aligned}$$

The corresponding starting procedure and underlying one-step methods are then the implicit Euler method and the following 2-stage Runge–Kutta method:

$$\begin{array}{ll}
S_h : & Y_n = y_n + \gamma h f(Y_n) \\
\Phi_h : & \text{Runge-Kutta method}
\end{array}
\quad
\begin{array}{c|cc}
\gamma & \gamma & \\
1 + \gamma & 1/2 + \gamma & 1/2 \\
\hline
& 1/2 + \gamma & 1/2 - \gamma
\end{array}$$

The method Φ_h is symmetric only for $\gamma = 0$, for $\gamma = 1/2$, and for $\gamma = -1/2$. This example demonstrates that the symmetry of the underlying one-step method strongly depends on the finishing procedure.

On the other hand, this example shows that the 2-stage Runge-Kutta method is symmetric in the sense of Definition 8.5 for all γ (because it is conjugate to the trapezoidal rule). It is not symmetric according to the definition of Chap. V.

A Useful Criterion for Symmetry. Definition 8.5 is rather impractical for verifying the symmetry of a given general linear method. We give here algebraic conditions for the coefficients A, B, C, D of a general linear method (8.1), which are sufficient for the method to be symmetric. We assume that the finishing procedure $y_{n+1} = F_h(Y_{n+1})$ is given by

$$y_{n+1} = \tilde{D}Y_{n+1} + h\tilde{B}f(V_{n+1}), \quad V_{n+1} = \tilde{C}Y_{n+1} + h\tilde{A}f(V_{n+1}), \quad (8.8)$$

in complete analogy to method (8.1).

Lemma 8.7 (Adjoint Method). *Let $Y_{n+1} = G_h(Y_n)$ be the general linear method given by A, B, C, D (with invertible D), $y_{n+1} = F_h(Y_{n+1})$ the finishing procedure given by $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}$, and denote by Φ_h its underlying one-step method. Then, the underlying one-step method of*

$$\begin{array}{ll}
G_h^* : & A^* = CD^{-1}B - A, \quad B^* = D^{-1}B, \quad C^* = CD^{-1}, \quad D^* = D^{-1} \\
F_h^* : & \tilde{A}^* = -\tilde{A}, \quad \tilde{B}^* = -\tilde{B}, \quad \tilde{C}^* = \tilde{C}, \quad \tilde{D}^* = \tilde{D}
\end{array}$$

is the adjoint method $\Phi_h^* = \Phi_{-h}^{-1}$ of Φ_h .

Proof. Substituting $h \leftrightarrow -h$ and $Y_{n+1} \leftrightarrow Y_n$ in (8.1) yields

$$U_{n+1} = CY_{n+1} - hAf(U_{n+1}), \quad Y_n = DY_{n+1} - hBf(U_{n+1}).$$

Extracting Y_{n+1} from the second relation and inserting it into the first gives

$$\begin{aligned}
U_{n+1} &= CD^{-1}Y_n + h(CD^{-1}B - A)f(U_{n+1}) \\
Y_{n+1} &= D^{-1}Y_n + hD^{-1}Bf(U_{n+1}),
\end{aligned}$$

which is exactly method G_h^* . The same replacements in the finishing procedure

$$V_{n+1} = \tilde{C}Y_n - h\tilde{A}f(V_{n+1}), \quad y_n = \tilde{D}Y_n - h\tilde{B}f(V_{n+1})$$

and in the diagram of Theorem 8.2 prove the statement. \square

Theorem 8.8. *If there exist an invertible matrix Q (satisfying $Qe = e$ with e given by (8.2)) and a permutation matrix P such that*

$$\begin{aligned} P^{-1}AP &= CD^{-1}B - A, & Q^{-1}BP &= D^{-1}B, \\ P^{-1}CQ &= CD^{-1}, & Q^{-1}DQ &= D^{-1}, \end{aligned} \quad (8.9)$$

then the general linear method (8.1) is symmetric.

Proof. We consider the change of variables $Y_n = Q\hat{Y}_n$, $U_n = P\hat{U}_n$ in the method (8.1). Since P is a permutation matrix, we have $f(PU) = Pf(U)$, so that the method becomes

$$P\hat{U}_{n+1} = CQ\hat{Y}_n + hAPf(\hat{U}_{n+1}), \quad Q\hat{Y}_{n+1} = DQ\hat{Y}_n + hBPf(\hat{U}_{n+1}).$$

The assumption (8.9) implies that this method is the same as the adjoint method of Lemma 8.7. Taking a finishing procedure F_h in such a way that $y_{n+1} = F_h(Q\hat{Y}_{n+1})$ is identical to the finishing procedure $y_{n+1} = F_h^*(\hat{Y}_{n+1})$ of the adjoint method (i.e., $\tilde{B} = 0$ and \tilde{D} such that $\tilde{D}Q = \tilde{D}$), we obtain $\Phi_h^* = \Phi_h$. This proves the statement. \square

The sufficient condition of Theorem 8.8 reduces to the known criteria for classical methods. Let us give some examples:

- For Runge–Kutta methods we have $D = (1)$, $B = b^T$ a row vector, and $C = \mathbb{1}$. With $Q = (1)$ and P the permutation matrix that inverts the elements of a vector, we get

$$b^T P = b^T, \quad PAP = \mathbb{1}b^T - A,$$

which is the same (V.2.4).

- Multistep methods in their form as general linear methods (Sect. XV.8) satisfy the condition of Theorem 8.8 if

$$\alpha_i = -\alpha_{k-i}, \quad \beta_i = \beta_{k-i}. \quad (8.10)$$

One can take for P and Q the permutation matrices (inverting the elements of a vector) of dimension $k+1$ and k , respectively.

XV.8.3 Growth Parameters

For a rigorous study of the long-time behaviour of general linear methods it is not sufficient to investigate smooth numerical solutions. One has to get bounds on the parasitic solution components, which are present when one considers the general linear method without any starting and finishing procedure. This is certainly difficult, as it is for multistep methods (1.1). We restrict here our analysis to the linearized parasitic modified equation.

The eigenvalues of the matrix D in (8.1) will play the role of the zeros of $\rho(\zeta)$ in (1.1). We denote them by $\zeta_1 = 1$ and ζ_2, \dots, ζ_k , and we assume that they are simple

and of modulus one. Motivated by the analysis for multistep methods we write the approximations Y_n as

$$Y_n = Y(nh) + \sum_{\ell \in \mathcal{I}^*} \zeta_\ell^n Z_\ell(nh) \quad (8.11)$$

with smooth functions $Y(t)$ and $Z_\ell(t)$. The index set \mathcal{I}^* has the same meaning as in Sect. XV.3.2. We insert (8.11) into (8.1) and compare coefficients of ζ_ℓ^n . This gives with $t = nh$

$$\begin{aligned} Y(t+h) &= DY(t) + hBf(CY(t)) + \mathcal{O}(h^2) \\ \zeta_\ell Z_\ell(t+h) &= DZ_\ell(t) + hBf'(CY(t))CZ_\ell(t) + \mathcal{O}(h^2). \end{aligned} \quad (8.12)$$

To get an amenable form of the modified equations we write the vectors $Y(t), Z_\ell(t)$ in the basis of eigenvectors of D , which we denote by $w_1 = e$ and w_2, \dots, w_k :

$$Y(t) = \sum_{j=1}^k y_j(t) w_j, \quad Z_\ell(t) = \sum_{j=1}^k z_{\ell,j}(t) w_j.$$

Inserted into (8.12) and expanded into a series of h yields

$$\dot{y}_1 = f(y_1) + \mathcal{O}(h),$$

and algebraic relations of the form $y_j(t) = \mathcal{O}(h)$ for $j \geq 2$. Similarly, we get algebraic relations for $z_{\ell,j}(t) = \mathcal{O}(h)$ if $j \neq \ell$, and the function $z_\ell(t) := z_{\ell,\ell}(t)$ satisfies

$$\dot{z}_\ell = \mu_\ell f'(y_1) z_\ell + \mathcal{O}(h) \quad \text{with} \quad \mu_\ell = \zeta_\ell^{-1} w_j^* B C w_j, \quad (8.13)$$

where w_j^* is the left eigenvector of D corresponding to the eigenvalue ζ_ℓ . This is in perfect analogy to the computations of Sect. XV.5.1.

This analysis can be extended straightforwardly to partitioned general linear methods, where different methods are applied to the components y and v of a partitioned differential equation. Unfortunately, we do not know of any results that would extend those of Sect. XV.6 to general linear methods.

XV.9 Exercises

1. Let $\zeta_1(z)$ be the principal root of the characteristic equation $\rho(\zeta) - z\sigma(\zeta) = 0$. Prove that for irreducible multistep methods the condition $\zeta_1(-z)\zeta_1(z) \equiv 1$ (in a neighbourhood of $z = 0$) is equivalent to the symmetry of the method.
2. (Lambert & Watson 1976). Prove that stable, symmetric linear multistep methods (1.8) for second order differential equations, for which the polynomial $\rho(\zeta)$ has only simple zeros (with the exception of $\zeta = 1$), has a non-vanishing interval of periodicity, i.e., the roots $\zeta_i(z)$ of $\rho(\zeta) - z^2\sigma(\zeta) = 0$ satisfy $|\zeta_i(iy)| = 1$ for sufficiently small real y .
Hint. Simple roots cannot leave the unit circle under small perturbations of y .

3. Consider a symmetric, s -stable multistep method (1.8). If it is irreducible (no common factors of $\rho(\zeta)$ and $\sigma(\zeta)$), then k is even. Hence $\rho(-1) \neq 0$.
4. Using Theorem XII.3.2, prove that the underlying one-step method of a strictly stable r th order linear multistep method has order r .
5. (Dahlquist 1959). Consider the linear problem $\dot{y} = \lambda y$ and apply a symmetric linear multistep method (1.1) as in Example 2.2. Prove that for $t = nh$ and $h \rightarrow 0$,

$$\zeta_j^n(\lambda h) \approx \zeta_j^n e^{\mu_j \lambda t},$$

where μ_j is the growth parameter.

6. Consider a general linear method (8.1). If there exist an invertible symmetric matrix G and a diagonal matrix Λ such that

$$M = \begin{pmatrix} D^T G D - G & D^T G B - C^T \Lambda \\ B^T G D - \Lambda C & B^T G B - A^T \Lambda - \Lambda A \end{pmatrix} = 0, \quad (9.1)$$

then the method is G -symplectic.

Hint. Adapt the proof of Burrage & Butcher for B -stability (see Hairer & Wanner (1996), page 358).

7. A Runge–Kutta method can be considered as a general linear method with $D = (1)$, $C = \mathbf{1}$. Prove that the condition (9.1) is equivalent to the symplecticity condition of Chap. VI.
8. Extend the definition of G -symplecticity to partitioned general linear methods, and prove that the condition

$$M = \begin{pmatrix} D^T G \hat{D} - G & D^T G \hat{B} - C^T \Lambda \\ B^T G \hat{D} - \Lambda \hat{C} & B^T G \hat{B} - A^T \Lambda - \Lambda \hat{A} \end{pmatrix} = 0 \quad (9.2)$$

implies that the method is G -symplectic.

9. Construct general linear methods of order $r > 2$, for which all growth parameters are positive. Find such methods, which have a smaller degree of implicitness than symmetric one-step methods of the same order.
10. Write a Maple program that checks the coefficients of Table 7.1. After defining `rho:=rho(z)`, use the instructions


```
> sigma := taylor(rho/(log(z)*log(z)), z=1, 8);
> factor(expand(convert(sigma, polynom)))
```
11. Construct partitioned general linear methods which are symmetric, explicit, of high order, and for which the matrices D and \hat{D} have distinct eigenvalues (with the exception of 1). Compared to multistep methods, smaller dimensions of the matrices D and \hat{D} are possible.

Bibliography

- R. Abraham & J.E. Marsden, *Foundations of Mechanics*, 2nd ed., Benjamin/Cummings Publishing Company, Reading, Massachusetts, 1978. [XIV.3]
- L. Abia & J.M. Sanz-Serna, *Partitioned Runge–Kutta methods for separable Hamiltonian problems*, Math. Comput. 60 (1993) 617–634. [VI.7], [IX.10]
- M.J. Ablowitz & J.F. Ladik, *A nonlinear difference scheme and inverse scattering*, Studies in Appl. Math. 55 (1976) 213–229. [VII.4]
- M.P. Allen & D.J. Tildesley, *Computer Simulation of Liquids*, Clarendon Press, Oxford, 1987. [I.4]
- H.C. Andersen, *Rattle: a “velocity” version of the Shake algorithm for molecular dynamics calculations*, J. Comput. Phys. 52 (1983) 24–34. [VII.1]
- V.I. Arnold, *Small denominators and problems of stability of motion in classical and celestial mechanics*, Russian Math. Surveys 18 (1963) 85–191. [I.1]
- V.I. Arnold, *Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l’hydrodynamique des fluides parfaites*, Ann. Inst. Fourier 16 (1966) 319–361. [VI.9]
- V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978, second edition 1989. [VI.1], [VII.2], [VII.5], [X.1], [X.7]
- V.I. Arnold, V.V. Kozlov & A.I. Neishtadt, *Mathematical Aspects of Classical and Celestial Mechanics*, Springer, Berlin, 1997. [X.1]
- U. Ascher & S. Reich, *On some difficulties in integrating highly oscillatory Hamiltonian systems*, in Computational Molecular Dynamics, Lect. Notes Comput. Sci. Eng. 4, Springer, Berlin, 1999, 281–296. [V.4]
- A. Aubry & P. Chartier, *Pseudo-symplectic Runge–Kutta methods*, BIT 38 (1998) 439–461. [X.7]
- H.F. Baker, *Alternants and continuous groups*, Proc. of London Math. Soc. 3 (1905) 24–47. [III.4]
- M.H. Beck, A. Jäckle, G.A. Worth & H.-D. Meyer, *The multiconfiguration time-dependent Hartree (MCTDH) method: A highly efficient algorithm for propagating wavepackets*, Phys. Reports 324 (2000) 1–105. [IV.9], [VII.6]
- G. Benettin, A.M. Cherubini & F. Fassò, *A changing-chart symplectic algorithm for rigid bodies and other Hamiltonian systems on manifolds*, SIAM J. Sci. Comput. 23 (2001) 1189–1203. [VII.4]
- G. Benettin, L. Galgani & A. Giorgilli, *Poincaré’s non-existence theorem and classical perturbation theory for nearly integrable Hamiltonian systems*, Advances in nonlinear dynamics and stochastic processes (Florence, 1985) World Sci. Publishing, Singapore, 1985, 1–22. [X.2]
- G. Benettin, L. Galgani & A. Giorgilli, *Realization of holonomic constraints and freezing of high frequency degrees of freedom in the light of classical perturbation theory. Part I*, Comm. Math. Phys. 113 (1987) 87–103. [XIII.6]

- G. Benettin, L. Galgani & A. Giorgilli, *Realization of holonomic constraints and freezing of high frequency degrees of freedom in the light of classical perturbation theory. II*, Commun. Math. Phys. 121 (1989) 557–601. [XIII.9]
- G. Benettin, L. Galgani, A. Giorgilli & J.-M. Strelcyn, *A proof of Kolmogorov's theorem on invariant tori using canonical transformations defined by the Lie method*, Il Nuovo Cimento 79B (1984) 201–223. [X.5]
- G. Benettin & A. Giorgilli, *On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms*, J. Statist. Phys. 74 (1994) 1117–1143. [IX.3], [IX.7], [IX.8]
- B.J. Berne, *Molecular dynamics in systems with multiple time scales: reference system propagator algorithms*, in Computational Molecular Dynamics: Challenges, Methods, Ideas (P. Deuflhard et al., eds.), Springer, Berlin 1999, 297–318. [XIII.1]
- Joh. Bernoulli, *Problème inverse des forces centrales, extrait de la réponse de Monsieur Bernoulli à Monsieur Herman*, Mém. de l'Acad. R. des Sciences de Paris (1710) p. 521, Opera Omnia I, p. 470–480. [I.2]
- M. Berry, *Histories of adiabatic quantum transitions*, Proc. Royal Soc. London A 429 (1990) 61–72. [XIV.1]
- V. Betz & S. Teufel, *Precise coupling terms in adiabatic quantum evolution*, Ann. Henri Poincaré 6 (2005) 217–246. [XIV.1]
- V. Betz & S. Teufel, *Precise coupling terms in adiabatic quantum evolution: the generic case*, Comm. Math. Phys., to appear (2005). [XIV.1]
- J.J. Biesiadecki & R.D. Skeel, *Dangers of multiple time step methods*, J. Comput. Phys. 109 (1993) 318–328. [I.4], [VIII.4], [XIII.1]
- G.D. Birkhoff, *Relativity and Modern Physics*, Harvard Univ. Press, Cambridge, Mass., 1923. [I.6]
- G.D. Birkhoff, *Dynamical Systems*, AMS, Providence, R.I., 1927. [X.2]
- S. Blanes, *High order numerical integrators for differential equations using composition and processing of low order methods*, Appl. Num. Math. (2001) 289–306. [V.3]
- S. Blanes & F. Casas, *On the necessity of negative coefficients for operator splitting schemes of order higher than two*, Appl. Num. Math. 54 (2005) 23–37. [III.3]
- S. Blanes, F. Casas & J. Ros, *Symplectic integrators with processing: a general study*, SIAM J. Sci. Comput. 21 (1999) 149–161. [V.3]
- S. Blanes, F. Casas & J. Ros, *Improved high order integrators based on the Magnus expansion*, BIT 40 (2000a) 434–450. [IV.7]
- S. Blanes, F. Casas & J. Ros, *Processing symplectic methods for near-integrable Hamiltonian systems*, Celestial Mech. Dynam. Astronom. 77 (2000b) 17–35. [V.3]
- S. Blanes & P.C. Moan, *Practical symplectic partitioned Runge–Kutta and Runge–Kutta–Nyström methods*, J. Comput. Appl. Math. 142 (2002) 313–330. [V.3]
- P.B. Bochev & C. Scovel, *On quadratic invariants and symplectic structure*, BIT 34 (1994) 337–345. [VI.4], [XV.4]
- N. Bogolioubov & I. Mitropolski, *Les Méthodes Asymptotiques en Théorie des Oscillations Non Linéaires*, Gauthier-Villars, Paris, 1962. [XII.2]
- N.N. Bogoliubov & Y.A. Mitropolsky, *Asymptotic Methods in the Theory of Non-Linear Oscillations*, Hindustan Publishing Corp., Delhi, 1961. [XII.1]
- J.F. Bonnans & J. Laurent-Varin, *Computation of order conditions for symplectic partitioned Runge–Kutta schemes with application to optimal control*, Numer. Math., to appear (2006). [VI.10]
- M. Born & V. Fock, *Beweis des Adiabatsatzes*, Zeitschr. f. Physik 51 (1928) 165–180. [XIV.1], [XIV.4]
- F. Bornemann, *Homogenization in Time of Singularly Perturbed Mechanical Systems*, Springer LNM 1687 (1998). [XIV.3]
- E. Bour, *L'intégration des équations différentielles de la mécanique analytique*, J. Math. Pures et Appliquées 20 (1855) 185–200. [X.1]

- K.E. Brenan, S.L. Campbell & L.R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Classics in Appl. Math., SIAM, Philadelphia, 1996. [IV.10]
- T.J. Bridges & S. Reich, *Computing Lyapunov exponents on a Stiefel manifold*, Physica D 156 (2001) 219–238. [IV.9], [IV.10]
- Ch. Brouder, *Runge–Kutta methods and renormalization*, Euro. Phys. J. C 12 (2000) 521–534. [III.1]
- Ch. Brouder, *Trees, Renormalization and Differential Equations*, BIT 44 (2004) 425–438. [III.1]
- C.J. Budd & M.D. Piggott, *Geometric integration and its applications*, Handbook of Numerical Analysis XI (2003) 35–139. [VIII.2]
- O. Buneman, *Time-reversible difference procedures*, J. Comput. Physics 1 (1967) 517–535. [V.1]
- C. Burrton & R. Scherer, *Gauss–Runge–Kutta–Nyström methods*, BIT 38 (1998) 12–21. [VI.10]
- K. Burrage & J.C. Butcher, *Stability criteria for implicit Runge–Kutta methods*, SIAM J. Numer. Anal. 16 (1979) 46–57. [VI.4]
- J.C. Butcher, *Coefficients for the study of Runge–Kutta integration processes*, J. Austral. Math. Soc. 3 (1963) 185–201. [II.1]
- J.C. Butcher, *Implicit Runge–Kutta processes*, Math. Comput. 18 (1964a) 50–64. [II.1]
- J.C. Butcher, *Integration processes based on Radau quadrature formulas*, Math. Comput. 18 (1964b) 233–244. [II.1]
- J.C. Butcher, *The effective order of Runge–Kutta methods*, in J.L. Morris, ed., Proceedings of Conference on the Numerical Solution of Differential Equations, Lecture Notes in Math. 109 (1969) 133–139. [V.3]
- J.C. Butcher, *An algebraic theory of integration methods*, Math. Comput. 26 (1972) 79–106. [III.1], [III.3]
- J.C. Butcher, *The Numerical Analysis of Ordinary Differential Equations. Runge–Kutta and General Linear Methods*, John Wiley & Sons, Chichester, 1987. [III.0], [III.1], [VI.7], [XV.8]
- J.C. Butcher, *Order and effective order*, Appl. Numer. Math. 28 (1998) 179–191. [V.3]
- J.C. Butcher & J.M. Sanz-Serna, *The number of conditions for a Runge–Kutta method to have effective order p* , Appl. Numer. Math. 22 (1996) 103–111. [III.1], [V.3]
- J.C. Butcher & G. Wanner, *Runge–Kutta methods: some historical notes*, Appl. Numer. Math. 22 (1996) 113–151. [III.1]
- M.P. Calvo, *High order starting iterates for implicit Runge–Kutta methods: an improvement for variable-step symplectic integrators*, IMA J. Numer. Anal. 22 (2002) 153–166. [VIII.6]
- M.P. Calvo & E. Hairer, *Accurate long-term integration of dynamical systems*, Appl. Numer. Math. 18 (1995a) 95–105. [X.3]
- M.P. Calvo & E. Hairer, *Further reduction in the number of independent order conditions for symplectic, explicit Partitioned Runge–Kutta and Runge–Kutta–Nyström methods*, Appl. Numer. Math. 18 (1995b) 107–114. [III.3]
- M.P. Calvo, A. Iserles & A. Zanna, *Numerical solution of isospectral flows*, Math. Comput. 66 (1997) 1461–1486. [IV.3]
- M.P. Calvo, A. Iserles & A. Zanna, *Conservative methods for the Toda lattice equations*, IMA J. Numer. Anal. 19 (1999) 509–523. [IV.3]
- M.P. Calvo, M.A. López-Marcos & J.M. Sanz-Serna, *Variable step implementation of geometric integrators*, Appl. Numer. Math. 28 (1998) 1–6. [VIII.2]
- M.P. Calvo, A. Murua & J.M. Sanz-Serna, *Modified equations for ODEs*, Contemporary Mathematics 172 (1994) 63–74. [IX.9]
- M.P. Calvo & J.M. Sanz-Serna, *Variable steps for symplectic integrators*, In: Numerical Analysis 1991 (Dundee, 1991), 34–48, Pitman Res. Notes Math. Ser. 260, 1992. [VIII.1]

- M.P. Calvo & J.M. Sanz-Serna, *The development of variable-step symplectic integrators, with application to the two-body problem*, SIAM J. Sci. Comput. 14 (1993) 936–952. [V.3], [X.3]
- M.P. Calvo & J.M. Sanz-Serna, *Canonical B-series*, Numer. Math. 67 (1994) 161–175. [VI.7]
- J. Candy & W. Rozmus, *A symplectic integration algorithm for separable Hamiltonian functions*, J. Comput. Phys. 92 (1991) 230–256. [II.5]
- B. Cano & A. Durán, *Analysis of variable-stepsize linear multistep methods with special emphasis on symmetric ones*, Math. Comp. 72 (2003) 1769–1801. [XV.7]
- B. Cano & A. Durán, *A technique to construct symmetric variable-stepsize linear multistep methods for second-order systems*, Math. Comp. 72 (2003) 1803–1816. [XV.7]
- B. Cano & J.M. Sanz-Serna, *Error growth in the numerical integration of periodic orbits by multistep methods, with application to reversible systems*, IMA J. Numer. Anal. 18 (1998) 57–75. [XV.5]
- R. Car & M. Parrinello, *Unified approach for molecular dynamics and density-functional theory*, Phys. Rev. Lett. 55 (1985) 2471–2474. [IV.9]
- J.R. Cash, *A class of implicit Runge–Kutta methods for the numerical integration of stiff ordinary differential equations*, J. Assoc. Comput. Mach. 22 (1975) 504–511. [II.3]
- A. Cayley, *On the theory of the analytic forms called trees*, Phil. Magazine XIII (1857) 172–176. [III.6]
- E. Celledoni & A. Iserles, *Methods for the approximation of the matrix exponential in a Lie-algebraic setting*, IMA J. Numer. Anal. 21 (2001) 463–488. [IV.8]
- R.P.K. Chan, *On symmetric Runge–Kutta methods of high order*, Computing 45 (1990) 301–309. [VI.10]
- P.J. Channell & J.C. Scovel, *Integrators for Lie–Poisson dynamical systems*, Phys. D 50 (1991) 80–88. [VII.5]
- P.J. Channell & J.C. Scovel, *Symplectic integration of Hamiltonian systems*, Nonlinearity 3 (1990) 231–259. [VI.5]
- S. Chaplygin, *A new case of motion of a heavy rigid body supported in one point* (Russian), Moscov Phys. Sect. 10, vol. 2 (1901). [X.1]
- P. Chartier, E. Faou & A. Murua, *An algebraic approach to invariant preserving integrators: the case of quadratic and Hamiltonian invariants*, Preprint, February 2005. [VI.7], [VI.8], [IX.9]
- M.T. Chu, *Matrix differential equations: a continuous realization process for linear algebra problems*, Nonlinear Anal. 18 (1992) 1125–1146. [IV.3]
- S. Cirilli, E. Hairer & B. Leimkuhler, *Asymptotic error analysis of the adaptive Verlet method*, BIT 39 (1999) 25–33. [VIII.3]
- A. Clebsch, *Ueber die simultane Integration linearer partieller Differentialgleichungen*, Crelle Journal f.d. reine u. angew. Math. 65 (1866) 257–268. [VII.3]
- D. Cohen, *Analysis and numerical treatment of highly oscillatory differential equations*, Doctoral Thesis, Univ. Geneva, 2004. [XIII.10]
- D. Cohen, *Conservation properties of numerical integrators for highly oscillatory Hamiltonian systems*, Report, 2005. To appear in IMA J. Numer. Anal. [XIII.10]
- D. Cohen, E. Hairer & Ch. Lubich, *Modulated Fourier expansions of highly oscillatory differential equations*, Found. Comput. Math. 3 (2003) 327–345. [XIII.6]
- D. Cohen, E. Hairer & Ch. Lubich, *Numerical energy conservation for multi-frequency oscillatory differential equations*, Report, 2004. To appear in BIT. [XIII.9]
- G.J. Cooper, *Stability of Runge–Kutta methods for trajectory problems*, IMA J. Numer. Anal. 7 (1987) 1–13. [IV.2]
- J.G. van der Corput, *Zur Methode der stationären Phase, I. Einfache Integrale*, Compos. Math. 1 (1934) 15–38. [XIV.4]
- M. Creutz & A. Gocksch, *Higher-order hybrid Monte Carlo algorithms*, Phys. Rev. Lett. 63 (1989) 9–12. [II.4]

- P.E. Crouch & R. Grossman, *Numerical integration of ordinary differential equations on manifolds*, J. Nonlinear Sci. 3 (1993) 1–33. [IV.8]
- M. Crouzeix, *Sur la B-stabilité des méthodes de Runge–Kutta*, Numer. Math. 32 (1979) 75–82. [VI.4]
- M. Crouzeix & J. Rappaz, *On Numerical Approximation in Bifurcation Theory*, Masson, Paris, 1989. [XIV.3]
- G. Dahlquist, *Convergence and stability in the numerical integration of ordinary differential equations*, Math. Scand. 4 (1956) 33–53. [XV.1]
- G. Dahlquist, *Stability and error bounds in the numerical integration of ordinary differential equations*, Trans. of the Royal Inst. of Techn. Stockholm, Sweden, Nr. 130 (1959) 87 pp. [XV.5], [XV.9]
- G. Dahlquist, *Error analysis for a class of methods for stiff nonlinear initial value problems*, Numerical Analysis, Dundee 1975, Lecture Notes in Math. 506 (1975) 60–74. [VI.8], [XV.4]
- G. Darboux, *Sur le problème de Pfaff*, extrait Bulletin des Sciences math. et astron. 2e série, vol. VI (1882); Gauthier-Villars, Paris, 1882. [VII.3]
- I. Degani & J. Schiff, *RCMS: Right correction Magnus series approach for integration of linear ordinary differential equations with highly oscillatory solution*, Report, Weizmann Inst. Science, Rehovot, 2003. [XIV.1]
- P. Deift, *Integrable Hamiltonian systems*, in P. Deift (ed.) et al., Dynamical systems and probabilistic methods in partial differential equations. AMS Lect. Appl. Math. 31 (1996) 103–138. [X.1]
- P. Deift, L.C. Li & C. Tomei, *Matrix factorizations and integrable systems*, Comm. Pure Appl. Math. 42 (1989) 443–521. [IV.3]
- P. Deift, L.C. Li & C. Tomei, *Symplectic aspects of some eigenvalue algorithms*, in A.S. Fokas & V.E. Zakharov (eds.), Important Developments in Soliton Theory, Springer 1993. [IV.3]
- P. Deift, T. Nanda & C. Tomei, *Ordinary differential equations and the symmetric eigenvalue problem*, SIAM J. Numer. Anal. 20 (1983) 1–22. [IV.3]
- P. Deuffhard, *A study of extrapolation methods based on multistep schemes without parasitic solutions*, Z. angew. Math. Phys. 30 (1979) 177–189. [XIII.1], [XIII.2]
- L. Dieci & T. Eirola, *On smooth decompositions of matrices*, SIAM J. Matrix Anal. Appl. 20 (1999) 800–819. [IV.9]
- L. Dieci, R.D. Russell & E.S. van Vleck, *Unitary integrators and applications to continuous orthonormalization techniques*, SIAM J. Numer. Anal. 31 (1994) 261–281. [IV.9]
- L. Dieci, R.D. Russell & E.S. van Vleck, *On the computation of Lyapunov exponents for continuous dynamical systems*, SIAM J. Numer. Anal. 34 (1997) 402–423. [IV.9], [IV.10]
- F. Diele, L. Lopez & R. Peluso, *The Cayley transform in the numerical solution of unitary differential systems*, Adv. Comput. Math. 8 (1998) 317–334. [IV.8]
- F. Diele, L. Lopez & T. Politi, *One step semi-explicit methods based on the Cayley transform for solving isospectral flows*, J. Comput. Appl. Math. 89 (1998) 219–223. [IV.3]
- P.A.M. Dirac, *Note on exchange phenomena in the Thomas atom*, Proc. Cambridge Phil. Soc. 26 (1930) 376–385. [IV.9], [VII.6]
- P.A.M. Dirac, *Generalized Hamiltonian dynamics*, Can. J. Math. 2 (1950) 129–148. [VII.7]
- V. Druskin & L. Knizhnerman, *Krylov subspace approximation of eigenpairs and matrix functions in exact and computer arithmetic*, Numer. Linear Algebra Appl. 2 (1995) 205–217. [XIII.1]
- A. Dullweber, B. Leimkuhler & R. McLachlan, *Symplectic splitting methods for rigid body molecular dynamics*, J. Chem. Phys. 107, No. 15 (1997) 5840–5851. [VII.4], [VII.5]
- W. E, *Analysis of the heterogeneous multiscale method for ordinary differential equations*, Comm. Math. Sci. 1 (2003) 423–436. [VIII.4]
- A. Edelman, T.A. Arias & S.T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl. 20 (1998) 303–353. [IV.9]

- B.L. Ehle, *On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems*, Research Report CSRR 2010 (1969), Dept. AACS, Univ. of Waterloo, Ontario, Canada. [II.1]
- E. Eich-Soellner & C. Führer, *Numerical Methods in Multibody Dynamics*, B. G. Teubner Stuttgart, 1998. [IV.4], [VII.1]
- T. Eirola, *Aspects of backward error analysis of numerical ODE's*, J. Comp. Appl. Math. 45 (1993), 65–73. [IX.1]
- T. Eirola & O. Nevanlinna, *What do multistep methods approximate?*, Numer. Math. 53 (1988) 559–569. [XV.2]
- T. Eirola & J.M. Sanz-Serna, *Conservation of integrals and symplectic structure in the integration of differential equations by multistep methods*, Numer. Math. 61 (1992) 281–290. [XV.4]
- L.H. Eliasson, *Absolutely convergent series expansions for quasi periodic motions*, Math. Phys. Electron. J. 2, No.4, Paper 4, 33 p. (1996). [X.2]
- K. Engø & S. Faltinsen, *Numerical integration of Lie–Poisson systems while preserving coadjoint orbits and energy*, SIAM J. Numer. Anal. 39 (2001) 128–145. [VII.5]
- B. Engquist & Y. Tsai, *Heterogeneous multiscale methods for stiff ordinary differential equations*, Math. Comp. 74 (2005) 1707–1742. [VIII.4]
- Ch. Engstler & Ch. Lubich, *Multirate extrapolation methods for differential equations with different time scales*, Computing 58 (1997) 173–185. [VIII.4]
- L. Euler, *Recherches sur la connoissance mécanique des corps*, Histoire de l'Acad. Royale de Berlin, Année MDCCLVIII, Tom. XIV, p. 131–153. Opera Omnia Ser. 2, Vol. 8, p. 178–199. [VII.5]
- L. Euler, *Du mouvement de rotation des corps solides autour d'un axe variable*, Hist. de l'Acad. Royale de Berlin, Tom. 14, Année MDCCLVIII, 154–193. Opera Omnia Ser. II, Vol. 8, 200–235. [IV.1]
- L. Euler, *Problème : un corps étant attiré en raison réciproque carrée des distances vers deux points fixes donnés, trouver les cas où la courbe décrite par ce corps sera algébrique*, Mémoires de l'Académie de Berlin for 1760, pub. 1767, 228–249. [X.1]
- L. Euler, *Theoria motus corporum solidorum seu rigidorum*, Rostochii et Gryphiswaldiae A.F. Röse, MDCCLXV. Opera Omnia Ser. 2, Vol. 3–4. [VII.5]
- L. Euler, *Institutionum Calculi Integralis*, Volumen Primum, Opera Omnia, Vol. XI. [I.1]
- E. Faou, E. Hairer & T.-L. Pham, *Energy conservation with non-symplectic methods: examples and counter-examples*, submitted for publication. [IX.9]
- E. Faou & Ch. Lubich, *A Poisson integrator for Gaussian wavepacket dynamics*, Report, 2004. To appear in Comp. Vis. Sci. [VII.4], [VII.6]
- F. Fassò, *Comparison of splitting algorithms for the rigid body*, J. Comput. Phys. 189 (2003) 527–538. [VII.5]
- K. Feng, *On difference schemes and symplectic geometry*, Proceedings of the 5-th Intern. Symposium on differential geometry & differential equations, August 1984, Beijing (1985) 42–58. [VI.3]
- K. Feng, *Difference schemes for Hamiltonian formalism and symplectic geometry*, J. Comp. Math. 4 (1986) 279–289. [VI.5]
- K. Feng, *Formal power series and numerical algorithms for dynamical systems*. In Proceedings of international conference on scientific computation, Hangzhou, China, Eds. Tony Chan & Zhong-Ci Shi, Series on Appl. Math. 1 (1991) 28–35. [IX.1]
- K. Feng, *Collected Works (II)*, National Defense Industry Press, Beijing, 1995. [XV.2]
- K. Feng & Z. Shang, *Volume-preserving algorithms for source-free dynamical systems*, Numer. Math. 71 (1995) 451–463. [IV.3]
- K. Feng, H.M. Wu, M.-Z. Qin & D.L. Wang, *Construction of canonical difference schemes for Hamiltonian formalism via generating functions*, J. Comp. Math. 7 (1989) 71–96. [VI.5]

- E. Fermi, J. Pasta & S. Ulam, *Studies of nonlinear problems*, Los Alamos Report No. LA-1940 (1955), later published in E. Fermi: *Collected Papers* (Chicago 1965), and *Lect. Appl. Math.* 15, 143 (1974). [I.5]
- B. Fiedler & J. Scheurle, *Discretization of homoclinic orbits, rapid forcing and “invisible” chaos*, *Mem. Amer. Math. Soc.* 119, no. 570, 1996. [IX.1]
- C.M. Field & F.W. Nijhoff, *A note on modified Hamiltonians for numerical integrations admitting an exact invariant*, *Nonlinearity* 16 (2003) 1673–1683. [IX.11]
- L.N.G. Filon, *On a quadrature formula for trigonometric integrals*, *Proc. Royal Soc. Edinburgh* 49 (1928) 38–47. [XIV.1]
- H. Flaschka, *The Toda lattice. II. Existence of integrals*, *Phys. Rev. B* 9 (1974) 1924–1925. [IV.3]
- J. Ford, *The Fermi–Pasta–Ulam problem: paradox turns discovery*, *Physics Reports* 213 (1992) 271–310. [I.5]
- E. Forest, *Canonical integrators as tracking codes*, *AIP Conference Proceedings* 184 (1989) 1106–1136. [II.4]
- E. Forest, *Sixth-order Lie group integrators*, *J. Comput. Physics* 99 (1992) 209–213. [V.3]
- E. Forest & R.D. Ruth, *Fourth-order symplectic integration*, *Phys. D* 43 (1990) 105–117. [II.5]
- J. Frenkel, *Wave Mechanics, Advanced General Theory*, Clarendon Press, Oxford, 1934. [IV.9], [VII.6]
- L. Galgani, A. Giorgilli, A. Martinoli & S. Vanzini, *On the problem of energy equipartition for large systems of the Fermi–Pasta–Ulam type: analytical and numerical estimates*, *Physica D* 59 (1992), 334–348. [I.5]
- M.J. Gander, *A non spiraling integrator for the Lotka Volterra equation*, *Il Volterriano* 4 (1994) 21–28. [VII.7]
- B. García-Archilla, J.M. Sanz-Serna & R.D. Skeel, *Long-time-step methods for oscillatory differential equations*, *SIAM J. Sci. Comput.* 20 (1999) 930–963. [VIII.4], [XIII.1], [XIII.2], [XIII.4]
- L.M. Garrido, *Generalized adiabatic invariance*, *J. Math. Phys.* 5 (1964) 355–362. [XIV.1]
- W. Gautschi, *Numerical integration of ordinary differential equations based on trigonometric polynomials*, *Numer. Math.* 3 (1961) 381–397. [XIII.1]
- Z. Ge & J.E. Marsden, *Lie–Poisson Hamilton–Jacobi theory and Lie–Poisson integrators*, *Phys. Lett. A* 133 (1988) 134–139. [VII.5], [IX.9]
- C.W. Gear & D.R. Wells, *Multirate linear multistep methods*, *BIT* 24 (1984) 484–502. [VIII.4]
- W. Gentzsch & A. Schlüter, *Über ein Einschnittverfahren mit zyklischer Schrittweitenänderung zur Lösung parabolischer Differentialgleichungen*, *ZAMM* 58 (1978), T415–T416. [II.4]
- S. Gill, *A process for the step-by-step integration of differential equations in an automatic digital computing machine*, *Proc. Cambridge Philos. Soc.* 47 (1951) 95–108. [III.1], [VIII.5]
- A. Giorgilli & U. Locatelli, *Kolmogorov theorem and classical perturbation theory*, *Z. Angew. Math. Phys.* 48 (1997) 220–261. [X.2]
- B. Gladman, M. Duncan & J. Candy, *Symplectic integrators for long-term integrations in celestial mechanics*, *Celestial Mechanics and Dynamical Astronomy* 52 (1991) 221–240. [VIII.1]
- D. Goldman & T.J. Kaper, *Nth-order operator splitting schemes and nonreversible systems*, *SIAM J. Numer. Anal.* 33 (1996) 349–367. [III.3]
- G.H. Golub & C.F. Van Loan, *Matrix Computations, 2nd edition*, John Hopkins Univ. Press, Baltimore and London, 1989. [IV.4]
- O. Gonzalez, *Time integration and discrete Hamiltonian systems*, *J. Nonlinear Sci.* 6 (1996) 449–467. [V.5]

- O. Gonzalez, D.J. Higham & A.M. Stuart, *Qualitative properties of modified equations*, IMA J. Numer. Anal. 19 (1999) 169–190. [IX.5]
- O. Gonzalez & J.C. Simo, *On the stability of symplectic and energy-momentum algorithms for nonlinear Hamiltonian systems with symmetry*, Comput. Methods Appl. Mech. Eng. 134 (1996) 197–222. [V.5]
- D.N. Goryachev, *On the motion of a heavy rigid body with an immobile point of support in the case $A = B = 4C$* (Russian), Moscov Math. Collect. 21 (1899) 431–438. [X.1]
- W.B. Gragg, *Repeated extrapolation to the limit in the numerical solution of ordinary differential equations*, Thesis, Univ. of California; see also SIAM J. Numer. Anal. 2 (1965) 384–403. [V.1]
- D.F. Griffiths & J.M. Sanz-Serna, *On the scope of the method of modified equations*, SIAM J. Sci. Stat. Comput. 7 (1986) 994–1008. [IX.1]
- V. Grimm & M. Hochbruck, *Error analysis of exponential integrators for oscillatory second-order differential equations*, Preprint, 2005. [XIII.4]
- W. Gröbner, *Die Lierihen und ihre Anwendungen*, VEB Deutscher Verlag der Wiss., Berlin 1960, 2nd ed. 1967. [III.5]
- H. Grubmüller, H. Heller, A. Windemuth & K. Schulten, *Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions*, Mol. Sim. 6 (1991) 121–142. [VIII.4], [XIII.1]
- A. Guillou & J.L. Soulé, *La résolution numérique des problèmes différentiels aux conditions initiales par des méthodes de collocation*, Rev. Française Informat. Recherche Opérationnelle 3 (1969) Ser. R-3, 17–44. [II.1]
- M. Günther & P. Rentrop, *Multirate ROW methods and latency of electric circuits*, Appl. Numer. Math. 13 (1993) 83–102. [VIII.4]
- F. Gustavson, *On constructing formal integrals of a Hamiltonian system near an equilibrium point*, Astron. J. 71 (1966) 670–686. [I.3]
- J. Hadamard, *Sur l'itération et les solutions asymptotiques des équations différentielles*, Bull. Soc. Math. France 29 (1901) 224–228. [XII.3]
- W.W. Hager, *Runge–Kutta methods in optimal control and the transformed adjoint system*, Numer. Math. 87 (2000) 247–282. [VI.10]
- E. Hairer, *Backward analysis of numerical integrators and symplectic methods*, Annals of Numerical Mathematics 1 (1994) 107–132. [VI.7]
- E. Hairer, *Variable time step integration with symplectic methods*, Appl. Numer. Math. 25 (1997) 219–227. [VIII.2]
- E. Hairer, *Backward error analysis for multistep methods*, Numer. Math. 84 (1999) 199–232. [IX.9], [XV.3]
- E. Hairer, *Symmetric projection methods for differential equations on manifolds*, BIT 40 (2000) 726–734. [V.4]
- E. Hairer, *Geometric integration of ordinary differential equations on manifolds*, BIT 41 (2001) 996–1007. [V.4]
- E. Hairer, *Global modified Hamiltonian for constrained symplectic integrators*, Numer. Math. 95 (2003) 325–336. [IX.5]
- E. Hairer & M. Hairer, *GniCodes – Matlab programs for geometric numerical integration*, In: Frontiers in numerical analysis (Durham, 2002), Springer Berlin, Universitext (2003), 199–240. [VIII.6]
- E. Hairer & P. Leone, *Order barriers for symplectic multi-value methods*. In: Numerical analysis 1997, Proc. of the 17th Dundee Biennial Conference, June 24–27, 1997, D.F. Griffiths, D.J. Higham & G.A. Watson (eds.), Pitman Research Notes in Mathematics Series 380 (1998), 133–149. [XV.4], [XV.8]
- E. Hairer & P. Leone, *Some properties of symplectic Runge–Kutta methods*, New Zealand J. of Math. 29 (2000) 169–175. [IV.2]

- E. Hairer & Ch. Lubich, *The life-span of backward error analysis for numerical integrators*, Numer. Math. 76 (1997), pp. 441–462. Erratum: <http://www.unige.ch/math/folks/hairer/> [IX.7], [X.5]
- E. Hairer & Ch. Lubich, *Invariant tori of dissipatively perturbed Hamiltonian systems under symplectic discretization*, Appl. Numer. Math. 29 (1999) 57–71. [XII.1], [XII.5]
- E. Hairer & Ch. Lubich, *Asymptotic expansions and backward analysis for numerical integrators*, Dynamics of Algorithms (Minneapolis, MN, 1997), IMA Vol. Math. Appl. 118, Springer, New York (2000) 91–106. [IX.1]
- E. Hairer & Ch. Lubich, *Long-time energy conservation of numerical methods for oscillatory differential equations*, SIAM J. Numer. Anal. 38 (2000a) 414–441. [XIII.1], [XIII.2], [XIII.5], [XIII.7]
- E. Hairer & Ch. Lubich, *Energy conservation by Störmer-type numerical integrators*, in: G.F. Griffiths, G.A. Watson (eds.), Numerical Analysis 1999, CRC Press LLC (2000b) 169–190. [XIII.8]
- E. Hairer & Ch. Lubich, *Symmetric multistep methods over long times*, Numer. Math. 97 (2004) 699–723. [XV.3], [XV.5], [XV.6]
- E. Hairer, Ch. Lubich & M. Roche, *The numerical solution of differential-algebraic systems by Runge–Kutta methods*, Lecture Notes in Math. 1409, Springer-Verlag, 1989. [VII.1]
- E. Hairer, Ch. Lubich & G. Wanner, *Geometric numerical integration illustrated by the Störmer–Verlet method*, Acta Numerica (2003) 399–450. [I.1]
- E. Hairer, S.P. Nørsett & G. Wanner, *Solving Ordinary Differential Equations I. Nonstiff Problems, 2nd edition*, Springer Series in Computational Mathematics 8, Springer Berlin, 1993. [II.1]
- E. Hairer & G. Söderlind, *Explicit, time reversible, adaptive step size control*, Submitted for publication, 2004. [VIII.3], [IX.6]
- E. Hairer & D. Stoffer, *Reversible long-term integration with variable stepsizes*, SIAM J. Sci. Comput. 18 (1997) 257–269. [VIII.3]
- E. Hairer & G. Wanner, *On the Butcher group and general multi-value methods*, Computing 13 (1974) 1–15. [III.1]
- E. Hairer & G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems, 2nd edition*, Springer Series in Computational Mathematics 14, Springer-Verlag Berlin, 1996. [II.1], [III.0], [IV.2], [IV.4], [IV.5], [IV.9], [IV.10], [VI.4], [VI.10], [VII.1], [VIII.6], [IX.5], [XIII.2], [XV.4], [XV.9]
- E. Hairer & G. Wanner, *Analysis by Its History, 2nd printing*, Undergraduate Texts in Mathematics, Springer-Verlag New York, 1997. [IX.7]
- M. Hall, jr., *A basis for free Lie rings and higher commutators in free groups*, Proc. Amer. Math. Soc. 1 (1950) 575–581. [III.3]
- Sir W.R. Hamilton, *On a general method in dynamics; by which the study of the motions of all free systems of attracting or repelling points is reduced to the search and differentiation of one central relation, or characteristic function*, Phil. Trans. Roy. Soc. Part II for 1834, 247–308; Math. Papers, Vol. II, 103–161. [VI.1], [VI.5]
- P.C. Hammer & J.W. Hollingsworth, *Trapezoidal methods of approximating solutions of differential equations*, MTAC 9 (1955) 92–96. [II.1]
- E.J. Haug, *Computer Aided Kinematics and Dynamics of Mechanical Systems, Volume I: Basic Methods*, Allyn & Bacon, Boston, 1989. [VII.5]
- F. Hausdorff, *Die symbolische Exponentialformel in der Gruppentheorie*, Berichte der Sächsischen Akad. der Wissensch. 58 (1906) 19–48. [III.4]
- A. Hayli, *Le problème des N corps dans un champ extérieur application à l'évolution dynamique des amas ouverts - I*, Bulletin Astronomique 2 (1967) 67–89. [VIII.4]
- R.B. Hayward, *On a Direct Method of estimating Velocities, Accelerations, and all similar Quantities with respect to Axes moveable in any Space, with Applications*, Cambridge Phil. Trans. vol. X (read 1856, publ. 1864) 1–20. [VII.5]

- E.J. Heller, *Time dependent approach to semiclassical dynamics*, J. Chem. Phys. 62 (1975) 1544–1555. [VII.6]
- E.J. Heller, *Time dependent variational approach to semiclassical dynamics*, J. Chem. Phys. 64 (1976) 63–73. [VII.6]
- M. Hénon & C. Heiles, *The applicability of the third integral of motion: some numerical experiments*, Astron. J. 69 (1964) 73–79. [I.3]
- J. Henrard, *The adiabatic invariant in classical mechanics*, Dynamics reported, New series. Vol. 2, Springer, Berlin (1993) 117–235. [XIV.1]
- P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley & Sons, Inc., New York 1962. [VIII.5]
- J. Hersch, *Contribution à la méthode aux différences*, Z. angew. Math. Phys. 9a (1958) 129–180. [XIII.1]
- K. Heun, *Neue Methode zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen*, Zeitschr. für Math. u. Phys. 45 (1900) 23–38. [II.1]
- D.J. Higham, *Time-stepping and preserving orthogonality*, BIT 37 (1997) 24–36. [IV.9]
- N.J. Higham, *The accuracy of floating point summation*, SIAM J. Sci. Comput. 14 (1993) 783–799. [VIII.5]
- M. Hochbruck & Ch. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal. 34 (1997) 1911–1925. [XIII.1]
- M. Hochbruck & Ch. Lubich, *A Gautschi-type method for oscillatory second-order differential equations*, Numer. Math. 83 (1999a) 403–426. [VIII.4], [XIII.1], [XIII.2], [XIII.4]
- M. Hochbruck & Ch. Lubich, *Exponential integrators for quantum-classical molecular dynamics*, BIT 39 (1999b) 620–645. [VIII.4], [XIV.1], [XIV.4]
- T. Holder, B. Leimkuhler & S. Reich, *Explicit variable step-size and time-reversible integration*, Appl. Numer. Math. 39 (2001) 367–377. [VIII.3]
- H. Hopf, *Über die Topologie der Gruppen-Mannigfaltigkeiten und ihre Verallgemeinerungen*, Ann. of Math. 42 (1941) 22–52. [III.1]
- W. Huang & B. Leimkuhler, *The adaptive Verlet method*, SIAM J. Sci. Comput. 18 (1997) 239–256. [VIII.2], [VIII.3]
- P. Hut, J. Makino & S. McMillan, *Building a better leapfrog*, Astrophys. J. 443 (1995) L93–L96. [VIII.3]
- K.J. In't Hout, *A new interpolation procedure for adapting Runge–Kutta methods to delay differential equations*, BIT 32 (1992) 634–649. [VIII.6]
- A. Iserles, *Solving linear ordinary differential equations by exponentials of iterated commutators*, Numer. Math. 45 (1984) 183–199. [II.4]
- A. Iserles, *On the global error of discretization methods for highly-oscillatory ordinary differential equations*, BIT 42 (2002) 561–599. [XIV.1]
- A. Iserles, *On the method of Neumann series for highly oscillatory equations*, BIT 44 (2004) 473–488. [XIV.1]
- A. Iserles, H.Z. Munthe-Kaas, S.P. Nørsett & A. Zanna, *Lie-group methods*, Acta Numerica (2000) 215–365. [IV.8]
- A. Iserles & S.P. Nørsett, *On the solution of linear differential equations in Lie groups*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci. 357 (1999) 983–1019. [IV.7], [IV.10]
- A. Iserles & S.P. Nørsett, *On the numerical quadrature of highly-oscillating integrals I: Fourier transforms*, IMA J. Numer. Anal. 24 (2004) 365–391. [XIV.1]
- T. Itoh & K. Abe, *Hamiltonian-conserving discrete canonical equations based on variational difference quotients*, J. Comput. Phys. 76 (1988) 85–102. [V.5]
- J.A. Izaguirre, S. Reich & R.D. Skeel, *Longer time steps for molecular dynamics*, J. Chem. Phys. 110 (1999) 9853–9864. [XIII.1], [XIV.4]
- C.G.J. Jacobi, *Über diejenigen Probleme der Mechanik, in welchen eine Kräftefunction existirt, und über die Theorie der Störungen*, manuscript from 1836 or 1837, published posthumely in *Werke*, vol. 5, 217–395. [VI.2]

- C.G.J. Jacobi, *Über die Reduktion der Integration der partiellen Differentialgleichungen erster Ordnung zwischen irgend einer Zahl Variablen auf die Integration eines einzigen Systemes gewöhnlicher Differentialgleichungen*, Crelle Journal f.d. reine u. angew. Math. 17 (1837) 97–162; K. Weierstrass, ed., C.G.J. Jacobi's Gesammelte Werke, vol. 4, pp. 57–127. [VI.5]
- C.G.J. Jacobi, *Lettre adressée à M. le Président de l'Académie des Sciences*, Liouville J. math. pures et appl. 5 (1840) 350–355; Werke, vol. 5, pp. 3–189. [IV.1]
- C.G.J. Jacobi, *Vorlesungen über Dynamik* (1842–43), Reimer, Berlin 1884. [VI.1], [VI.5], [VI.6], [VI.10]
- C.G.J. Jacobi, *Nova methodus, aequationes differentiales partiales primi ordinis inter numerum variabilium quemcunque propositas integrandi*, published posthumly in Crelle Journal f.d. reine u. angew. Math. 60 (1861) 1–181; Werke, vol. 5, pp. 3–189. [III.5], [VII.2], [VII.3]
- T. Jahnke, *Numerische Verfahren für fast adiabatische Quantendynamik*, Doctoral Thesis, Univ. Tübingen, 2003. [XIV.3]
- T. Jahnke, *Long-time-step integrators for almost-adiabatic quantum dynamics*, SIAM J. Sci. Comput. 25 (2004a) 2145–2164. [XIV.1]
- T. Jahnke, *A long-time-step method for quantum-classical molecular dynamics*, Report, 2004b. [XIV.3]
- T. Jahnke & Ch. Lubich, *Numerical integrators for quantum dynamics close to the adiabatic limit*, Numer. Math. 94 (2003), 289–314. [XIV.1]
- L. Jay, *Collocation methods for differential-algebraic equations of index 3*, Numer. Math. 65 (1993) 407–421. [VII.1]
- L. Jay, *Runge–Kutta type methods for index three differential-algebraic equations with applications to Hamiltonian systems*, Thesis No. 2658, 1994, Univ. Genève. [VII.1]
- L. Jay, *Symplectic partitioned Runge–Kutta methods for constrained Hamiltonian systems*, SIAM J. Numer. Anal. 33 (1996) 368–387. [II.2], [VII.1]
- L. Jay, *Specialized Runge–Kutta methods for index 2 differential algebraic equations*, Math. Comp. (2005), to appear. [IV.9]
- R. Jost, *Winkel- und Wirkungsvariable für allgemeine mechanische Systeme*, Helv. Phys. Acta 41 (1968) 965–968. [X.1]
- A. Joye & C.-E. Pfister, *Superadiabatic evolution and adiabatic transition probability between two nondegenerate levels isolated in the spectrum*, J. Math. Phys. 34 (1993) 454–479. [XIV.1]
- W. Kahan, *Further remarks on reducing truncation errors*, Comm. ACM 8 (1965) 40. [VIII.5]
- W. Kahan & R.-C. Li, *Composition constants for raising the orders of unconventional schemes for ordinary differential equations*, Math. Comput. 66 (1997) 1089–1099. [V.3], [V.6]
- B. Karasözen, *Poisson integrators*, Math. Comp. Modelling 40 (2004) 1225–1244. [VII.4]
- T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer, Berlin, 1980. [VII.6]
- J. Kepler, *Astronomia nova αλτιολογητός seu Physica celestis, traditua commentariis de motibus stellae Martis, ex observationibus G. V. Tychonis Brahe*, Prague 1609. [I.2]
- H. Kinoshita, H. Yoshida & H. Nakai, *Symplectic integrators and their application to dynamical astronomy*, Celest. Mech. & Dynam. Astr. 50 (1991) 59–71. [V.3]
- U. Kirchgraber, *Multi-step methods are essentially one-step methods*, Numer. Math. 48 (1986) 85–90. [XV.2]
- U. Kirchgraber, F. Lasagni, K. Nipp & D. Stoffer, *On the application of invariant manifold theory, in particular to numerical analysis*, Internat. Ser. Numer. Math. 97, Birkhäuser, Basel, 1991, 189–197. [XII.3]
- U. Kirchgraber & E. Stiefel, *Methoden der analytischen Störungsrechnung und ihre Anwendungen*, Teubner, Stuttgart, 1978. [XII.4]

- F. Klein, *Elementarmathematik vom höheren Standpunkte aus. Teil I: Arithmetik, Algebra, Analysis*, ausgearbeitet von E. Hellinger, Teubner, Leipzig, 1908; Vierte Auflage, Die Grundlehren der mathematischen Wissenschaften, Band 14 Springer-Verlag, Berlin, 1933, reprinted 1968. [VII.5]
- F. Klein & A. Sommerfeld, *Theorie des Kreisels*, Leipzig 1897. [VII.5]
- O. Koch & Ch. Lubich, *Dynamical low rank approximation*, Preprint, 2005. [IV.9]
- A.N. Kolmogorov, *On conservation of conditionally periodic motions under small perturbations of the Hamiltonian*, Dokl. Akad. Nauk SSSR 98 (1954) 527–530. [X.2], [X.5]
- A.N. Kolmogorov, *General theory of dynamical systems and classical mechanics*, Proc. Int. Congr. Math. Amsterdam 1954, Vol. 1, 315–333. [X.2], [X.5]
- P.-V. Koseleff, *Exhaustive search of symplectic integrators using computer algebra*, Integration algorithms and classical mechanics, Fields Inst. Commun. 10 (1996) 103–120. [V.3]
- S. Kovalevskaya (Kowalevski), *Sur le problème de la rotation d'un corps solide autour d'un point fixe*, Acta Math. 12 (1889) 177–232. [X.1]
- V.V. Kozlov, *Integrability and non-integrability in Hamiltonian mechanics*, Uspekhi Mat. Nauk 38 (1983) 3–67. [X.1]
- D. Kreimer, *On the Hopf algebra structure of perturbative quantum field theory*, Adv. Theor. Math. Phys. 2 (1998) 303–334. [III.1]
- N.M. Krylov & N.N. Bogoliubov, *Application des méthodes de la mécanique non linéaire à la théorie des oscillations stationnaires*, Edition de l'Académie des Sciences de la R.S.S. d'Ukraine, 1934. [XII.4]
- W. Kutta, *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*, Zeitschr. für Math. u. Phys. 46 (1901) 435–453. [II.1]
- R.A. LaBudde & D. Greenspan, *Discrete mechanics – a general treatment*, J. Comput. Phys. 15 (1974) 134–167. [V.5]
- R.A. LaBudde & D. Greenspan, *Energy and momentum conserving methods of arbitrary order for the numerical integration of equations of motion. Parts I and II*, Numer. Math. 25 (1976) 323–346 and 26 (1976) 1–26. [V.5]
- M.P. Laburta, *Starting algorithms for IRK methods*, J. Comput. Appl. Math. 83 (1997) 269–288. [VIII.6]
- M.P. Laburta, *Construction of starting algorithms for the RK-Gauss methods*, J. Comput. Appl. Math. 90 (1998) 239–261. [VIII.6]
- J.-L. Lagrange, *Applications de la méthode exposée dans le mémoire précédent à la solution de différents problèmes de dynamique*, 1760, Oeuvres Vol. 1, 365–468. [VI.1], [VI.2]
- J.L. Lagrange, *Recherches sur le mouvement d'un corps qui est attiré vers deux centres fixes* (1766), Œuvres, tome II, Gauthier-Villars, Paris 1868, 67–124. [X.1]
- J.-L. Lagrange, *Mécanique analytique*, Paris 1788. [VI.1]
- J.D. Lambert & I.A. Watson, *Symmetric multistep methods for periodic initial value problems*, J. Inst. Maths. Applics. 18 (1976) 189–202. [XV.1], [XV.9]
- C. Lanczos, *The Variational Principles of Mechanics*, University of Toronto Press, Toronto, 1949. (Fourth edition 1970). [VI.6]
- P.S. Laplace, *Traité de mécanique céleste II*, 1799, see Œuvres I, p. 183. [I.6]
- F.M. Lasagni, *Canonical Runge–Kutta methods*, ZAMP 39 (1988) 952–953. [VI.4], [VI.5], [VI.7]
- J.D. Lawson, *Generalized Runge–Kutta processes for stable systems with large Lipschitz constants*, SIAM J. Numer. Anal. 4 (1967) 372–380. [XIV.1]
- P.D. Lax, *Integrals of nonlinear equations of evolution and solitary waves*, Commun. Pure Appl. Math. 21 (1968) 467–490. [IV.3]
- B. Leimkuhler & S. Reich, *Symplectic integration of constrained Hamiltonian systems*, Math. Comp. 63 (1994) 589–605. [VII.1]
- B. Leimkuhler & S. Reich, *A reversible averaging integrator for multiple time-scale dynamics*, J. Comput. Phys. 171 (2001) 95–114. [VIII.4]

- B. Leimkuhler & S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge Monographs on Applied and Computational Mathematics **14**, Cambridge University Press, Cambridge, 2004. [VI.3]
- B.J. Leimkuhler & R.D. Skeel, *Symplectic numerical integrators in constrained Hamiltonian systems*, J. Comput. Phys. **112** (1994) 117–125. [VII.1]
- A. Lenard, *Adiabatic invariance to all orders*, Ann. Phys. **6** (1959) 261–276. [XIV.1]
- P. Leone, *Symplecticity and Symmetry of General Integration Methods*, Thèse, Section de Mathématiques, Université de Genève, 2000. [VI.8]
- T. Levi-Civita, *Sur la résolution qualitative du problème restreint des trois corps*, Acta Math. **30** (1906) 305–327. [VIII.2]
- T. Levi-Civita, *Sur la régularisation du problème des trois corps*, Acta Math. **42** (1920) 99–144. [VIII.2]
- D. Lewis & J.C. Simo, *Conserving algorithms for the dynamics of Hamiltonian systems on Lie groups*, J. Nonlinear Sci. **4** (1994) 253–299. [IV.8], [V.5]
- D. Lewis & J.C. Simo, *Conserving algorithms for the N-dimensional rigid body*, Fields Inst. Com. **10** (1996) 121–139. [V.5]
- S. Lie, *Zur Theorie der Transformationsgruppen*, Christ. Forh. Aar. 1888, Nr. 13, 6 pages, Christiania 1888; Gesammelte Abh. vol. 5, p. 553–557. [VII.2], [VII.3]
- J. Liouville, *Note à l'occasion du mémoire précédent (de M. E. Bour)*, J. Math. Pures et Appliquées **20** (1855) 201–202. [X.1]
- L. Lopez & T. Politi, *Applications of the Cayley approach in the numerical solution of matrix differential systems on quadratic groups*, Appl. Numer. Math. **36** (2001) 35–55. [IV.8]
- M.A. López-Marcos, J.M. Sanz-Serna & R.D. Skeel, *Cheap enhancement of symplectic integrators*, Numerical analysis 1995 (Dundee), Pitman Res. Notes Math. Ser. **344**, Longman, Harlow, 1996, 107–122. [V.3]
- K. Lorenz, T. Jahnke & Ch. Lubich, *Adiabatic integrators for highly oscillatory second order linear differential equations with time-varying eigendecomposition*, BIT **45** (2005) 91–115. [XIV.1], [XIV.2]
- A.J. Lotka, *The Elements of Physical Biology*, Williams & Wilkins, Baltimore, 1925. Reprinted 1956 under the title *Elements of mathematical biology* by Dover, New York. [I.1]
- Ch. Lubich, *Integration of stiff mechanical systems by Runge-Kutta methods*, Z. Angew. Math. Phys. **44** (1993) 1022–1053. [XIV.3]
- Ch. Lubich, *On dynamics and bifurcations of nonlinear evolution equations under numerical discretization*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems (B. Fiedler, ed.), Springer, Berlin, 2001, 469–500. [XII.3]
- Ch. Lubich, *A variational splitting integrator for quantum molecular dynamics*, Appl. Numer. Math. **48** (2004) 355–368. [VII.4]
- Ch. Lubich, *On variational approximations in quantum molecular dynamics*, Math. Comp. **74** (2005) 765–779. [VII.6]
- R. MacKay, *Some aspects of the dynamics of Hamiltonian systems*, in: D.S. Broomhead & A. Iserles, eds., *The Dynamics of Numerics and the Numerics of Dynamics*, Clarendon Press, Oxford, 1992, 137–193. [VI.6]
- S. Maeda, *Canonical structure and symmetries for discrete systems*, Math. Japonica **25** (1980) 405–420. [VI.6]
- S. Maeda, *Lagrangian formulation of discrete systems and concept of difference space*, Math. Japonica **27** (1982) 345–356. [VI.6]
- W. Magnus, *On the exponential solution of differential equations for a linear operator*, Comm. Pure Appl. Math. **VII** (1954) 649–673. [IV.7]
- G. Marchuk, *Some applications of splitting-up methods to the solution of mathematical physics problems*, Aplikace Matematiky **13** (1968) 103–132. [II.5]
- J.E. Marsden, S. Pekarsky & S. Shkoller, *Discrete Euler-Poincaré and Lie-Poisson equations*, Nonlinearity **12** (1999) 1647–1662. [VII.5]

- J.E. Marsden & T.S. Ratiu, *Introduction to Mechanics and Symmetry. A Basic Exposition of Classical Mechanical Systems*, Second edition, Texts in Applied Mathematics 17, Springer-Verlag, New York, 1999. [IV.1]
- J.E. Marsden & M. West, *Discrete mechanics and variational integrators*, Acta Numerica 10 (2001) 1–158. [VI.6]
- A.D. McLachlan, *A variational solution of the time-dependent Schrodinger equation*, Mol. Phys. 8 (1964) 39–44. [VII.6]
- R.I. McLachlan, *Explicit Lie-Poisson integration and the Euler equations*, Phys. Rev. Lett. 71 (1993) 3043–3046. [VII.4], [VII.5]
- R.I. McLachlan, *On the numerical integration of ordinary differential equations by symmetric composition methods*, SIAM J. Sci. Comput. 16 (1995) 151–168. [II.4], [II.5], [III.3], [V.3], [V.6]
- R.I. McLachlan, *Composition methods in the presence of small parameters*, BIT 35 (1995b) 258–268. [V.3]
- R.I. McLachlan, *More on symplectic integrators*, in *Integration Algorithms and Classical Mechanics* 10, J.E. Marsden, G.W. Patrick & W.F. Shadwick, eds., Amer. Math. Soc., Providence, R.I. (1996) 141–149. [V.3]
- R.I. McLachlan, *Featured review of Geometric Numerical Integration by E. Hairer, C. Lubich, and G. Wanner*, SIAM Review 45 (2003) 817–821. [VII.5]
- R.I. McLachlan & P. Atela, *The accuracy of symplectic integrators*, Nonlinearity 5 (1992) 541–562. [V.3]
- R.I. McLachlan & G.R.W. Quispel, *Splitting methods*, Acta Numerica 11 (2002) 341–434. [VII.4]
- R.I. McLachlan, G.R.W. Quispel & N. Robidoux, *Geometric integration using discrete gradients*, Philos. Trans. R. Soc. Lond., Ser. A, 357 (1999) 1021–1045. [V.5]
- R.I. McLachlan & C. Scovel, *Equivariant constrained symplectic integration*, J. Nonlinear Sci. 5 (1995) 233–256. [VII.5]
- R.I. McLachlan & A. Zanna, *The discrete Moser–Veselov algorithm for the free rigid body, revisited*, Found. Comput. Math. 5 (2005) 87–123. [VII.5], [IX.11]
- R.J.Y. McLeod & J.M. Sanz-Serna, *Geometrically derived difference formulae for the numerical integration of trajectory problems*, IMA J. Numer. Anal. 2 (1982) 357–370. [VIII.2]
- V.L. Mehrmann, *The Autonomous Linear Quadratic Control Problem. Theory and Numerical Solution*, Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1991. [IV.9]
- R.H. Merson, *An operational method for the study of integration processes*, Proc. Symp. Data Processing, Weapons Research Establishment, Salisbury, Australia (1957) 110–1 to 110–25. [III.1]
- A. Messiah, *Quantum Mechanics*, Dover Publ., 1999 (reprint of the two-volume edition published by Wiley, 1961–1962). [VII.6]
- S. Miesbach & H.J. Pesch, *Symplectic phase flow approximation for the numerical integration of canonical systems*, Numer. Math. 61 (1992) 501–521. [VI.5]
- P.C. Moan, *On rigorous modified equations for discretizations of ODEs*, Report, 2005. [IX.7]
- O. Möller, *Quasi double-precision in floating point addition*, BIT 5 (1965) 37–50 and 251–255. [VIII.5]
- A. Morbidelli & A. Giorgilli, *Superexponential stability of KAM Tori*, J. Stat. Phys. 78 (1995) 1607–1617. [X.2]
- J. Moser, *Review MR 20-4066*, Math. Rev., 1959. [X.5]
- J. Moser, *On invariant curves of area-preserving mappings of an annulus*, Nachr. Akad. Wiss. Göttingen, II. Math.-Phys. Kl. 1962, 1–20. [X.5]
- J. Moser, *Lectures on Hamiltonian systems*, Mem. Am. Math. Soc. 81 (1968) 1–60. [IX.3]
- J. Moser, *Stable and Random Motions in Dynamical Systems*, Annals of Mathematics Studies. No. 77. Princeton University Press, 1973. [XI.2]

- J. Moser, *Finitely many mass points on the line under the influence of an exponential potential — an integrable system*, Dyn. Syst., Theor. Appl., Battelle Seattle 1974 Renc., Lect. Notes Phys. 38 (1975) 467–497. [X.1]
- J. Moser, *Is the solar system stable?*, Mathematical Intelligencer 1 (1978) 65–71. [X.0]
- J. Moser & A.P. Veselov, *Discrete versions of some classical integrable systems and factorization of matrix polynomials*, Commun. Math. Phys. 139 (1991) 217–243. [VII.5]
- H. Munthe-Kaas, *Lie Butcher theory for Runge–Kutta methods*, BIT 35 (1995) 572–587. [IV.8]
- H. Munthe-Kaas, *Runge–Kutta methods on Lie groups*, BIT 38 (1998) 92–111. [IV.8]
- H. Munthe-Kaas, *High order Runge–Kutta methods on manifolds*, J. Appl. Num. Maths. 29 (1999) 115–127. [IV.8]
- H. Munthe-Kaas & B. Owren, *Computations in a free Lie algebra*, Phil. Trans. Royal Soc. A 357 (1999) 957–981. [IV.7]
- A. Murua, *Métodos simplécticos desarrollables en P-series*, Doctoral Thesis, Univ. Valladolid, 1994. [IX.3]
- A. Murua, *On order conditions for partitioned symplectic methods*, SIAM J. Numer. Anal. 34 (1997) 2204–2211. [IX.11]
- A. Murua, *Formal series and numerical integrators, Part I: Systems of ODEs and symplectic integrators*, Appl. Numer. Math. 29 (1999) 221–251. [IX.11]
- A. Murua & J.M. Sanz-Serna, *Order conditions for numerical integrators obtained by composing simpler integrators*, Philos. Trans. Royal Soc. London, ser. A 357 (1999) 1079–1100. [III.1], [III.3], [V.3]
- A.I. Neishtadt, *The separation of motions in systems with rapidly rotating phase*, J. Appl. Math. Mech. 48 (1984) 133–139. [XIV.2]
- N.N. Nekhoroshev, *An exponential estimate of the time of stability of nearly-integrable Hamiltonian systems*, Russ. Math. Surveys 32 (1977) 1–65. [X.2], [X.4]
- N.N. Nekhoroshev, *An exponential estimate of the time of stability of nearly-integrable Hamiltonian systems. II.* (Russian), Tr. Semin. Im. I.G. Petrovskogo 5 (1979) 5–50. [X.4]
- G. Nenciu, *Linear adiabatic theory. Exponential estimates*, Commun. Math. Phys. 152 (1993) 479–496. [XIV.1]
- P. Nettesheim & S. Reich, *Symplectic multiple-time-stepping integrators for quantum-classical molecular dynamics*, in P. Deuflhard et al. (eds.), Computational Molecular Dynamics: Challenges, Methods, Ideas, Springer, Berlin 1999, 412–420. [VIII.4]
- I. Newton, *Philosophiae Naturalis Principia Mathematica*, Londini anno MDCLXXXVII, 1687. [I.2], [VI.1], [X.1]
- I. Newton, *Second edition of the Principia*, 1713. [I.2], [X.1]
- K. Nipp & D. Stoffer, *Attractive invariant manifolds for maps: existence, smoothness and continuous dependence on the map*, Research Report No. 92–11, SAM, ETH Zürich, 1992. [XII.3]
- K. Nipp & D. Stoffer, *Invariant manifolds and global error estimates of numerical integration schemes applied to stiff systems of singular perturbation type. I: RK-methods*, Numer. Math. 70 (1995) 245–257. [XII.3]
- K. Nipp & D. Stoffer, *Invariant manifolds and global error estimates of numerical integration schemes applied to stiff systems of singular perturbation type. II: Linear multistep methods*, Numer. Math. 74 (1996) 305–323. [XII.3]
- E. Noether, *Invariante Variationsprobleme*, Nachr. Akad. Wiss. Göttingen, Math.-Phys. Kl. (1918) 235–257. [VI.6]
- E.J. Nyström, *Ueber die numerische Integration von Differentialgleichungen*, Acta Soc. Sci. Fenn. 50 (1925) 1–54. [II.2]
- E. Oja, *Neural networks, principal components, and subspaces*, Int. J. Neural Syst. 1 (1989) 61–68. [IV.9]
- D. Okunbor & R.D. Skeel, *Explicit canonical methods for Hamiltonian systems*, Math. Comp. 59 (1992) 439–455. [VI.4]

- D.I. Okunbor & R.D. Skeel, *Canonical Runge–Kutta–Nyström methods of orders five and six*, J. Comp. Appl. Math. 51 (1994) 375–382. [V.3]
- F.W.J. Olver, *Asymptotics and Special Functions*, Academic Press, 1974. [XIV.4]
- P.J. Olver, *Applications of Lie Groups to Differential Equations*, Graduate Texts in Mathematics 107, Springer-Verlag, New York, 1986. [IV.6]
- B. Owren & A. Marthinsen, *Runge–Kutta methods adapted to manifolds and based on rigid frames*, BIT 39 (1999) 116–142. [IV.8]
- B. Owren & A. Marthinsen, *Integration methods based on canonical coordinates of the second kind*, Numer. Math. 87 (2001) 763–790. [IV.8]
- A.M. Perelomov, *Selected topics on classical integrable systems*, Troisième cycle de la physique, expanded version of lectures delivered in May 1995. [VII.2]
- O. Perron, *Über Stabilität und asymptotisches Verhalten der Lösungen eines Systems endlicher Differenzengleichungen*, J. Reine Angew. Math. 161 (1929) 41–64. [XII.3]
- A.D. Perry & S. Wiggins, *KAM tori are very sticky: Rigorous lower bounds on the time to move away from an invariant Lagrangian torus with linear flow*, Physica D 71 (1994) 102–121. [X.2]
- H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste, Tome I*, Gauthier-Villars, Paris, 1892. [VI.1], [X.1], [X.2]
- H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste, Tome II*, Gauthier-Villars, Paris, 1893. [VI.1], [X.2]
- H. Poincaré, *Les Méthodes Nouvelles de la Mécanique Céleste. Tome III*, Gauthiers-Villars, Paris, 1899. [VI.1], [VI.2]
- L. Poinso, *Théorie nouvelle de la rotation des corps*, Paris 1834. [VII.5]
- S.D. Poisson, *Sur la variation des constantes arbitraires dans les questions de mécanique*, J. de l'Ecole Polytechnique vol. 8, 15e cahier (1809) 266–344. [VII.2]
- B. van der Pol, *Forced oscillations in a system with non-linear resistance*, Phil. Mag. 3, (1927), 65–80; *Papers* vol. I, 361–376. [XII.4]
- J. Pöschel, *Nekhoroshev estimates for quasi-convex Hamiltonian systems*, Math. Z. 213 (1993) 187–216. [X.2]
- F.A. Potra & W.C. Rheinboldt, *On the numerical solution of Euler–Lagrange equations*, Mech. Struct. & Mech. 19 (1991) 1–18. [IV.5]
- M.-Z. Qin & W.-J. Zhu, *Volume-preserving schemes and numerical experiments*, Comput. Math. Appl. 26 (1993) 33–42. [VI.9]
- G.D. Quinlan, *Resonances and instabilities in symmetric multistep methods*, Report, 1999, available on <http://xxx.lanl.gov/abs/astro-ph/9901136> [XV.7]
- G.D. Quinlan & S. Tremaine, *Symmetric multistep methods for the numerical integration of planetary orbits*, Astron. J. 100 (1990) 1694–1700. [XV.1], [XV.7]
- G.R.W. Quispel, *Volume-preserving integrators*, Phys. Lett. A 206 (1995) 26–30. [VI.9]
- S. Reich, *Symplectic integration of constrained Hamiltonian systems by Runge–Kutta methods*, Techn. Report 93-13 (1993), Dept. Comput. Sci., Univ. of British Columbia. [VII.1]
- S. Reich, *Numerical integration of the generalized Euler equations*, Techn. Report 93-20 (1993), Dept. Comput. Sci., Univ. of British Columbia. [VII.4]
- S. Reich, *Momentum conserving symplectic integrators*, Phys. D 76 (1994) 375–383. [VII.5]
- S. Reich, *Symplectic integration of constrained Hamiltonian systems by composition methods*, SIAM J. Numer. Anal. 33 (1996a) 475–491. [VII.1], [IX.5]
- S. Reich, *Enhancing energy conserving methods*, BIT 36 (1996b) 122–134. [V.5]
- S. Reich, *Backward error analysis for numerical integrators*, SIAM J. Numer. Anal. 36 (1999) 1549–1570. [VIII.2], [IX.5], [IX.7]
- J.R. Rice, *Split Runge–Kutta method for simultaneous equations*, J. Res. Nat. Bur. Standards 64B (1960) 151–170. [VIII.4]
- H. Rubin & P. Ungar, *Motion under a strong constraining force*, Comm. Pure Appl. Math. 10 (1957) 65–87. [XIV.3]

- C. Runge, *Ueber die numerische Auflösung von Differentialgleichungen*, Math. Ann. 46 (1895) 167–178. [II.1]
- H. Rüssmann, *On optimal estimates for the solutions of linear partial differential equations of first order with constant coefficients on the torus*, Dyn. Syst., Theor. Appl., Battelle Seattle 1974 Renc., Lect. Notes Phys. 38 (1975) 598–624. [X.4]
- H. Rüssmann, *On optimal estimates for the solutions of linear difference equations on the circle*, Celest. Mech. 14 (1976) 33–37. [X.4]
- R.D. Ruth, *A canonical integration technique*, IEEE Trans. Nuclear Science NS-30 (1983) 2669–2671. [II.5], [VI.1], [VI.3], [IX.1]
- J.-P. Ryckaert, G. Cicciotti & H.J.C. Berendsen, *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes*, J. Comput. Phys. 23 (1977) 327–341. [VII.1], [XIII.1]
- P. Saha & S. Tremaine, *Symplectic integrators for solar system dynamics*, Astron. J. 104 (1992) 1633–1640. [V.3]
- S. Saito, H. Sugiura & T. Mitsui, *Butcher's simplifying assumption for symplectic integrators*, BIT 32 (1992) 345–349. [IV.10]
- J. Sand, *Methods for starting iteration schemes for implicit Runge–Kutta formulae*, Computational ordinary differential equations (London, 1989), Inst. Math. Appl. Conf. Ser. New Ser., 39, Oxford Univ. Press, New York, 1992, 115–126. [VIII.6]
- J.M. Sanz-Serna, *Runge–Kutta schemes for Hamiltonian systems*, BIT 28 (1988) 877–883. [VI.4]
- J.M. Sanz-Serna, *Symplectic integrators for Hamiltonian problems: an overview*, Acta Numerica 1 (1992) 243–286. [IX.1]
- J.M. Sanz-Serna, *An unconventional symplectic integrator of W. Kahan*, Appl. Numer. Math. 16 (1994) 245–250. [VII.4]
- J.M. Sanz-Serna & L. Abia, *Order conditions for canonical Runge–Kutta schemes*, SIAM J. Numer. Anal. 28 (1991) 1081–1096. [IV.10]
- J.M. Sanz-Serna & M.P. Calvo, *Numerical Hamiltonian Problems*, Chapman & Hall, London, 1994. [VI.3], [VIII.6]
- R. Scherer, *A note on Radau and Lobatto formulae for O.D.E:s*, BIT 17 (1977) 235–238. [II.3]
- T. Schlick, *Some failures and successes of long-timestep approaches to biomolecular simulations*, in Computational Molecular Dynamics: Challenges, Methods, Ideas (P. Deuffhard et al., eds.), Springer, Berlin 1999, 227–262. [XIII.1]
- M.B. Sevryuk, *Reversible systems*, Lecture Notes in Mathematics, 1211. Springer-Verlag, 1986. [XI.0]
- L.F. Shampine, *Conservation laws and the numerical solution of ODEs*, Comp. Maths. Appls. 12B (1986) 1287–1296. [IV.1]
- Z. Shang, *Generating functions for volume-preserving mappings and Hamilton–Jacobi equations for source-free dynamical systems*, Sci. China Ser. A 37 (1994a) 1172–1188. [VI.9]
- Z. Shang, *Construction of volume-preserving difference schemes for source-free systems via generating functions*, J. Comput. Math. 12 (1994b) 265–272. [VI.9]
- Z. Shang, *KAM theorem of symplectic algorithms for Hamiltonian systems*, Numer. Math. 83 (1999) 477–496. [X.6]
- Z. Shang, *Resonant and Diophantine step sizes in computing invariant tori of Hamiltonian systems*, Nonlinearity 13 (2000) 299–308. [X.6]
- Q. Sheng, *Solving linear partial differential equations by exponential splitting*, IMA J. Numer. Anal. 9 (1989) 199–212. [III.3]
- C.L. Siegel & J.K. Moser, *Lectures on Celestial Mechanics*, Grundlehren d. math. Wiss. vol. 187, Springer-Verlag 1971; First German edition: C.L. Siegel, *Vorlesungen über Himmelsmechanik*, Grundlehren vol. 85, Springer-Verlag, 1956. [VI.1], [VI.5], [VI.6]
- J.C. Simo & N. Tarnow, *The discrete energy-momentum method. Conserving algorithms for nonlinear elastodynamics*, Z. Angew. Math. Phys. 43 (1992) 757–792. [V.5]

- J.C. Simo, N. Tarnow & K.K. Wong, *Exact energy-momentum conserving algorithms and symplectic schemes for nonlinear dynamics*, Comput. Methods Appl. Mech. Eng. 100 (1992) 63–116. [V.5]
- H.D. Simon & H. Zha, *Low rank matrix approximation using the Lanczos bidiagonalization process with applications*, SIAM J. Sci. Comput. 21 (2000) 2257–2274. [IV.9]
- R.D. Skeel & C.W. Gear, *Does variable step size ruin a symplectic integrator?*, Physica D60 (1992) 311–313. [VIII.2]
- M. Sofroniou & G. Spaletta, *Derivation of symmetric composition constants for symmetric integrators*, J. of Optimization Methods and Software (2004) to appear. [V.3]
- A. Sommerfeld, *Mechanics* (Lectures on Theoretical Physics, vol. I), first German ed. 1942, English transl. by M.O. Stern, Acad. Press. [VII.5]
- S. Sternberg, *Celestial Mechanics*, Benjamin, New York, 1969. [X.0]
- E. Stiefel, *Richtungsfelder und Fernparallelismus in n -dimensionalen Mannigfaltigkeiten*, Comment. Math. Helv. 8 (1935) 305–353. [IV.9]
- H.J. Stetter, *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag, Berlin, 1973. [II.3], [II.4], [V.1], [V.2]
- D. Stoffer, *On reversible and canonical integration methods*, SAM-Report No. 88-05, ETH-Zürich, 1988. [V.1]
- D. Stoffer, *Variable steps for reversible integration methods*, Computing 55 (1995) 1–22. [VIII.2], [VIII.3]
- D. Stoffer, *General linear methods: connection to one step methods and invariant curves*, Numer. Math. 64 (1993) 395–407. [XV.2]
- D. Stoffer, *On the qualitative behaviour of symplectic integrators. III: Perturbed integrable systems*, J. Math. Anal. Appl. 217 (1998) 521–545. [XII.4]
- C. Störmer, *Sur les trajectoires des corpuscules électrisés*, Arch. sci. phys. nat., Genève, vol. 24 (1907) 5–18, 113–158, 221–247. [I.1]
- G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal. 5 (1968) 506–517. [II.5]
- W.B. Strett, D.J. Tildesley & G. Saville, *Multiple time step methods in molecular dynamics*, Mol. Phys. 35 (1978) 639–648. [VIII.4]
- A.M. Stuart & A.R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, 1996. [XII.3]
- G. Sun, *Construction of high order symplectic Runge–Kutta Methods*, J. Comput. Math. 11 (1993a) 250–260. [IV.2]
- G. Sun, *Symplectic partitioned Runge–Kutta methods*, J. Comput. Math. 11 (1993b) 365–372. [II.2], [IV.2]
- G. Sun, *A simple way constructing symplectic Runge–Kutta methods*, J. Comput. Math. 18 (2000) 61–68. [VI.10]
- K.F. Sundman, *Mémoire sur le problème des trois corps*, Acta Math. 36 (1912) 105–179. [VIII.2]
- Y.B. Suris, *On the conservation of the symplectic structure in the numerical solution of Hamiltonian systems* (in Russian), In: Numerical Solution of Ordinary Differential Equations, ed. S.S. Filippov, Keldysh Institute of Applied Mathematics, USSR Academy of Sciences, Moscow, 1988, 148–160. [VI.4]
- Y.B. Suris, *The canonicity of mappings generated by Runge–Kutta type methods when integrating the systems $\ddot{x} = -\partial U/\partial x$* , Zh. Vychisl. Mat. i Mat. Fiz. 29, 202–211 (in Russian); same as U.S.S.R. Comput. Maths. Phys. 29 (1989) 138–144. [VI.4]
- Y.B. Suris, *Hamiltonian methods of Runge–Kutta type and their variational interpretation* (in Russian), Math. Model. 2 (1990) 78–87. [VI.6]
- Y.B. Suris, *Partitioned Runge–Kutta methods as phase volume preserving integrators*, Phys. Lett. A 220 (1996) 63–69. [VI.9]
- Y.B. Suris, *Integrable discretizations for lattice systems: local equations of motion and their Hamiltonian properties*, Rev. Math. Phys. 11 (1999) 727–822. [VII.2]

- Y.B. Suris, *The Problem of Integrable Discretization: Hamiltonian Approach*, Progress in Mathematics 219, Birkhäuser, Basel, 2003. [X.3]
- G.J. Sussman & J. Wisdom, *Chaotic evolution of the solar system*, Science 257 (1992) 56–62. [I.2]
- M. Suzuki, *Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations*, Phys. Lett. A 146 (1990) 319–323. [II.4], [II.5]
- M. Suzuki, *General theory of fractal path integrals with applications to many-body theories and statistical physics*, J. Math. Phys. 32 (1991) 400–407. [III.3]
- M. Suzuki, *General theory of higher-order decomposition of exponential operators and symplectic integrators*, Phys. Lett. A 165 (1992) 387–395. [II.5], [V.6]
- M. Suzuki, *Quantum Monte Carlo methods and general decomposition theory of exponential operators and symplectic integrators*, Physica A 205 (1994) 65–79. [V.3]
- M. Suzuki & K. Umeno, *Higher-order decomposition theory of exponential operators and its applications to QMC and nonlinear dynamics*, In: Computer Simulation Studies in Condensed-Matter Physics VI, Landau, Mon, Schüttler (eds.), Springer Proceedings in Physics 76 (1993) 74–86. [V.3]
- W.W. Symes, *The QR algorithm and scattering for the finite nonperiodic Toda lattice*, Physica D 4 (1982) 275–280. [IV.3]
- F. Takens, *Motion under the influence of a strong constraining force*, Global theory of dynamical systems, Proc. int. Conf., Evanston/Ill. 1979, Springer LNM 819 (1980) 425–445. [XIV.3]
- Y.-F. Tang, *The symplecticity of multi-step methods*, Computers Math. Applic. 25 (1993) 83–90. [XV.4]
- Y.-F. Tang, *Formal energy of a symplectic scheme for Hamiltonian systems and its applications (I)*, Computers Math. Applic. 27 (1994) 31–39. [IX.3]
- Y.-F. Tang, V.M. Pérez-García & L. Vázquez, *Symplectic methods for the Ablowitz–Ladik model*, Appl. Math. Comput. 82 (1997) 17–38. [VII.4]
- B. Thaller, *Visual Quantum Mechanics. Selected topics with computer-generated animations of quantum-mechanical phenomena*. Springer-TELOS, New York, 2000. [VII.6]
- W. Thirring, *Lehrbuch der Mathematischen Physik I*, Springer-Verlag, 1977. [X.5]
- M. Toda, *Waves in nonlinear lattice*, Progr. Theor. Phys. Suppl. 45 (1970) 174–200. [X.1]
- J. Touma & J. Wisdom, *Lie–Poisson integrators for rigid body dynamics in the solar system*, Astron. J. 107 (1994) 1189–1202. [VII.5]
- H.F. Trotter, *On the product of semi-groups of operators*, Proc. Am. Math. Soc. 10 (1959) 545–551. [II.5]
- M. Tuckerman, B.J. Berne & G.J. Martyna, *Reversible multiple time scale molecular dynamics*, J. Chem. Phys. 97 (1992) 1990–2001. [VIII.4], [XIII.1]
- V.S. Varadarajan, *Lie Groups, Lie Algebras and Their Representations*, Prentice-Hall, Englewood Cliffs, New Jersey, 1974 [III.4], [IV.6], [IV.8]
- L. Verlet, *Computer “experiments” on classical fluids. I. Thermodynamical properties of Lennard–Jones molecules*, Physical Review 159 (1967) 98–103. [I.1], [XIII.1]
- A.P. Veselov, *Integrable systems with discrete time, and difference operators*, Funktsional. Anal. i Prilozhen. 22 (1988) 1–13, 96; transl. in Funct. Anal. Appl. 22 (1988) 83–93. [VI.6]
- A.P. Veselov, *Integrable maps*, Russ. Math. Surv. 46 (1991) 1–51. [VI.6]
- R. de Vogelaere, *Methods of integration which preserve the contact transformation property of the Hamiltonian equations*, Report No. 4, Dept. Math., Univ. of Notre Dame, Notre Dame, Ind. (1956) [I.1], [VI.3]
- V. Volterra, *Variazioni e fluttuazioni del numero d’individui in specie animali conviventi*, Mem. R. Comitato talassografico italiano, CXXXI, 1927; Opere 5, p. 1–111. [I.1]

- J. Waldvogel & F. Spirig, *Chaotic motion in Hill's lunar problem*, In: A.E. Roy and B.A. Steves, eds., *From Newton to Chaos: Modern Techniques for Understanding and Coping with Chaos in N-Body Dynamical Systems* (NATO Adv. Sci. Inst. Ser. B Phys., 336, Plenum Press, New York, 1995). [VIII.2]
- G. Wanner, *Runge–Kutta-methods with expansion in even powers of h* , Computing 11 (1973) 81–85. [III.3], [V.2]
- R.A. Wehage & E.J. Haug, *Generalized coordinate partitioning for dimension reduction in analysis of constrained dynamic systems*, J. Mechanical Design 104 (1982) 247–255. [IV.5]
- J.M. Wendlandt & J.E. Marsden, *Mechanical integrators derived from a discrete variational principle*, Physica D 106 (1997) 223–246. [VI.6]
- H. Weyl, *The Classical Groups*, Princeton Univ. Press, Princeton, 1939. [VI.2]
- H. Weyl, *The method of orthogonal projection in potential theory*, Duke Math. J. 7 (1940) 411–444. [VI.9]
- J.H. Wilkinson, *Error analysis of floating-point computation*, Numer. Math. 2 (1960) 319–340. [IX.0]
- J. Wisdom & M. Holman, *Symplectic maps for the N-body problem*, Astron. J. 102 (1991) 1528–1538. [V.3]
- J. Wisdom, M. Holman & J. Touma, *Symplectic correctors*, in *Integration Algorithms and Classical Mechanics* 10, J.E. Marsden, G.W. Patrick & W.F. Shadwick, eds., Amer. Math. Soc., Providence, R.I. (1996) 217–244. [V.3]
- K. Wright, *Some relationships between implicit Runge–Kutta, collocation and Lanczos τ methods, and their stability properties*, BIT 10 (1970) 217–227. [II.1]
- K. Wright, *Differential equations for the analytic singular value decomposition of a matrix*, Numer. Math. 63 (1992) 283–295. [IV.9]
- W.Y. Yan, U. Helmke & J.B. Moore, *Global analysis of Oja's flow for neural networks*, IEEE Trans. Neural Netw. 5 (1994) 674–683. [IV.9]
- H. Yoshida, *Construction of higher order symplectic integrators*, Phys. Lett. A 150 (1990) 262–268. [II.4], [II.5], [III.4], [III.5], [V.3]
- H. Yoshida, *Recent progress in the theory and application of symplectic integrators*, Celestial Mech. Dynam. Astronom. 56 (1993) 27–43. [IX.1], [IX.4], [IX.8]
- A. Zanna, *Collocation and relaxed collocation for the Fer and the Magnus expansions*, SIAM J. Numer. Anal. 36 (1999) 1145–1182. [IV.7], [IV.10]
- A. Zanna, K. Engø & H.Z. Munthe-Kaas, *Adjoint and selfadjoint Lie-group methods*, BIT 41 (2001) 395–421. [V.4], [V.6]
- K. Zare & V. Szebehely, *Time transformations in the extended phase-space*, Celestial Mechanics 11 (1975) 469–482. [VIII.2]
- C. Zener, *Non-adiabatic crossing of energy levels*, Proc. Royal Soc. London, Ser. A 137 (1932) 696–702. [XIV.1]
- S.L. Ziglin, *The ABC-flow is not integrable for $A = B$* , Funktsional. Anal. i Prilozhen. 30 (1996) 80–81; transl. in Funct. Anal. Appl. 30 (1996) 137–138. [VI.9]

Index

- ABC flow 228
- Abel–Liouville–Jacobi–Ostrogradskii identity 105, 228
- Ablowitz–Ladik model 273
- action integral 205
- action-angle variables 397
- adaptive Verlet method 309
- adiabatic integrator 547
- adiabatic invariants 531, 533, 545, 562
- adiabatic transformations 531, 532
- adjoint method 42, 145, 342, 613
 - of collocation method 146
 - of Runge–Kutta method 147
 - quadratic invariants 176
- adjoint operator 83
- angular momentum 9, 98, 100, 101, 276, 591, 601
- area preservation 5, 6, 183, 184, 188
- Argon crystal 19
- Arnold–Liouville theorem 397
- attractive invariant manifold 460, 574, 610
- attractive invariant torus 464
 - of numerical integrator 467
- averaged forces 319
- averaging
 - basic scheme 458
 - perturbation series 459
- averaging principle 456
- avoided crossing 535, 563

- B-series 51, 56, 57, 212, 223, 575
 - composition 61
 - symplectic 217, 219
- backward error analysis 337, 576
 - formal 337
 - rigorous 360
- BCH formula 83, 84, 348
 - symmetric 86
- Bernoulli numbers 84, 122
- bi-coloured trees 66
- B_∞ -series 72

- Birkhoff normalization
 - Hamiltonian 412
 - reversible 447
- $B(p)$ 32
- Butcher group 64, 372
- Butcher product 75, 212

- canonical 186
 - equations of motion 181
 - form 267
 - Poisson structure 254
 - transformation 186
- canonical coordinates of a Lie group
 - first kind 129
 - second kind 129
- Casimir function 257, 267, 283
- Cayley transform 107, 128
- central field 392, 400, 438, 440
- characteristic lines 262
- Choleski decomposition 154
- coadjoint orbit 287
- collocation methods 27, 30, 122
 - discontinuous 35, 247
 - symmetric 146
- collocation polynomial 30
- commutator 118
 - matrix 83
- compensated summation 323
- complete systems 263
- completely integrable 393
- composition
 - of B-series 61
 - of Runge–Kutta methods 59
- composition methods 43, 45, 50, 92, 105, 190, 333
 - ρ -compatibility 145
 - local error 150
 - of order 2 150
 - of order 4 152, 155
 - of order 6 153, 156
 - of order 8 157

- of order 10 158
- order conditions 71, 75, 80
- symmetric 149
- symmetric-symmetric 154
- with symmetric method 154
- conditionally periodic flow 399
- configuration manifold 239
- conjugate momenta 181
- conjugate symplecticity 222, 225, 592
- conservation
 - of area 5, 183
 - of energy 98, 172, 366, 484, 512, 600
 - of linear invariants 99
 - of mass 98
 - of momentum 172, 600
 - of quadratic invariants 101, 102, 212, 214, 216
 - of volume 227
- conserved quantity 97
- consistent initial values 238
- constant direction of projection 165
- constrained Hamiltonian systems 239, 258
- constrained mechanical systems 237
- continuous output 326
- coordinates
 - generalized 180
- cotangent bundle 240
- $C(q)$ 32
- Crouch-Grossman methods 124
 - order conditions 124
- d'Alembert principle 259
- Darboux–Lie theorem 261, 265, 266, 272
- degrees of freedom 5
- diagonally implicit Runge–Kutta methods
 - symmetric 147
- differential equations 2
 - Hamilton–Jacobi 200
 - Hamiltonian 4, 180
 - highly oscillatory 21
 - modified 337
 - on Lie groups 118
 - on manifolds 115, 239
 - partial, linear 262
 - reversible 143
 - second order 7, 41, 216, 332
- differential equations on manifolds
 - ρ -compatibility 145
- differential form 186
- differential-algebraic equations 140, 237
- diophantine frequencies 406
- Dirac–Frenkel variational principle 138, 259, 295
- DIRK methods
 - symmetric 147
- discontinuous collocation 35, 247
- discrete Euler–Lagrange equations 206
- discrete Lagrangian 206
- discrete momenta 206
- discrete-gradient methods 171, 174
- dissipative systems 455
- distinguished functions 266
- divergence-free vector fields 227
- eccentricity 9
- effective order
 - of composition methods 158
- EI 150
- elementary differentials 52, 53, 66
- elementary Hamiltonian 373, 384
- elementary weights 55
- energy
 - oscillatory 479, 484, 505, 510, 517, 524
 - total 182, 479, 484, 510, 524
- energy conservation 366, 379, 510, 524, 600
- energy exchange 483, 490, 494
- energy-momentum methods 171
 - for N -body systems 173
- equistage approximation 329
- error analysis
 - backward 337
- error growth
 - linear 413, 414, 448
 - of rounding errors 324
- Euler equations 275, 277, 279
- Euler method
 - –Lie 126
 - explicit 3
 - implicit 3
 - symplectic 4, 48, 189, 230, 242, 270
- Euler parameters 281
- Euler–Lagrange equations 181, 205, 237
 - discrete 206
- explicit symmetric methods 148
- exponential map 120
- Fermi–Pasta–Ulam problem 21, 479
- filter function 481
- first integrals 5, 97, 211, 375
 - long-time near-preservation 413, 448
 - quadratic 212, 591
- fixed-point iteration 330
- flow 2
 - discrete 3
 - exact 2, 49, 200

- isospectral 107
- numerical 3, 49
- Poisson 261, 265
- frequencies 399
- diophantine 406
- Frobenius norm 132
- G-symplectic 587
- Gauss methods 34, 101, 333
 - symmetric 147
 - symplectic 192
- Gaussian wavepacket 296
- Gautschi-type methods 473, 477
- general linear methods 609
 - strictly stable 609
 - symmetric 611
 - weakly stable 610
- generalized coordinate partitioning 117
- generating functions 195, 197, 201, 204, 288, 344
 - for partitioned RK methods 199
 - for Runge–Kutta methods 198
- geometrical numerical algebra 131
- $GL(n)$, general linear group 119
- $\mathfrak{gl}(n)$, Lie algebra of $n \times n$ matrices 119
- Grassmann manifold 131, 135
- growth parameter 592, 614
- Hénon–Heiles problem 380
- Hall set 78
- Hamilton’s principle 204, 205
- Hamilton–Jacobi equation 200, 391
- Hamiltonian 4, 181, 257
 - elementary 373, 384
 - global 186
 - local 185, 234
 - modified 343, 375
- Hamiltonian perturbation theory 389, 404
 - basic scheme 405
 - Birkhoff normalization 412
 - KAM theory 410, 423
 - perturbation series 406
- Hamiltonian systems 4, 180
 - constrained 239, 258, 289
 - integrable 390
 - non-canonical 237
 - perturbed integrable 404
- harmonic oscillator
 - varying frequency 546
- heavy top 283
- Hénon–Heiles model 15
- Hopf algebra 65
- IE 150
- implementation 303, 325
- implicit midpoint rule 3, 34, 190, 192, 223, 270
 - averaged 537
 - symmetry 145
 - symplecticity 190
- impulse method 317, 475, 550
 - mollified 476
- index reduction 239, 241
- inertia ellipsoid 275
- integrability lemma 186
- integrable systems 601
 - Hamiltonian 390
 - reversible 437
- invariant manifold 574
 - attractive 460, 574, 610
- invariant torus 397, 423
 - long-time near-preservation 422, 451
 - of numerical integrator 433, 453, 467
 - of reversible map 451
 - of symplectic map 431
 - weakly attractive 464
- invariants 2, 5, 97
 - adiabatic 531, 533, 545, 562
 - linear 99
 - polynomial 105
 - quadratic 101
 - weak 109
- involution
 - first integrals in 391
- irreducible
 - Runge–Kutta methods 220
- isospectral flow 107, 403
- isospectral methods 107
- iteration
 - fixed-point 330
 - Newton-type 331
- Jacobi identity 118, 255
- KAM theory
 - Hamiltonian 410, 423
 - reversible 445
 - reversible near-identity map 451
 - symplectic near-identity map 431
- KAM torus
 - sticky 412
- Kepler problem 8, 25, 46, 111, 150, 234, 416, 603
 - perturbed 12, 26, 304
- Kepler’s second law 9
- kernel
 - of processing methods 158

- kinetic energy 180, 237
- Kolmogorov's iteration 410
- Kolmogorov's theorem 423
- Lagrange equations 181
- Lagrange multipliers 111, 132, 237, 279
- Lax pair 403
- leap-frog method 7
- left-invariant 289
- Legendre transform 181
 - discrete 206
- Leibniz' rule 255
- Lennard–Jones potential 19
- Lie algebras 118, 286
- Lie bracket 89, 118, 261
 - differential operators 89
- Lie derivative 87, 348, 362
 - of B-series 370
 - of P-series 382
- Lie group methods 123, 351
 - symmetric 169
- Lie groups 118
 - quadratic 128
- Lie midpoint rule 127
- Lie operator 261
- Lie–Euler method 126
- Lie–Poisson reduction 289
- Lie–Poisson systems 274, 286
- Lie–Trotter splitting 47
- Lindstedt–Poincaré series 406
- linear error growth 12, 413, 414, 448, 601
- linear multistep methods
 - weakly stable 575
- linear stability 23
- Liouville lemma 392
- Liouville's theorem 227
- Lobatto IIIA - IIIB pair 102, 192, 210, 247, 352, 386
- Lobatto IIIA methods 34, 377
 - symmetric 147
- Lobatto IIIA–IIIB pair 40
- Lobatto IIIB methods 37, 377, 449
 - symmetric 147
- Lobatto IIIS 235
- Lobatto quadrature 247
- local coordinates 113
 - existence of numerical solution 167
 - symmetric methods 166
- local error 29
 - of composition methods 150, 176
- long-time behaviour
 - symmetric integrators 437, 455
 - symplectic integrators 389, 455
- long-time energy conservation 366
- Lorenz problem 176
- Lotka–Volterra problem 1, 24, 175, 257, 270, 271, 273, 340
- low-rank approximation 137
- Lyapunov exponents 131
- Magnus series 121
- manifold of rank k matrices 131
- manifolds 109, 114, 239, 267
 - symmetric methods 161
 - symplectic 258
- Marchuk splitting 47
- matrix commutator 83
- matrix exponential 120
- matrix Lie groups 118
- mechanical systems 555
 - constrained 237, 258
- merging product 75
- methods based on local coordinates 166
- methods on manifolds 97, 350
 - symmetric 161
- midpoint rule 123
 - explicit 569, 580
 - implicit 3, 34, 190, 192, 223, 270
 - Lie 127
 - modified 171
- modified differential equation 337
 - B-series 369
 - constrained Hamiltonian system 352
 - first integrals 351
 - Lie group methods 351
 - Lie–Poisson integrators 354
 - methods on manifolds 350
 - P-series 381
 - perturbed differential equation 466
 - Poisson integrators 347
 - reversible methods 343
 - splitting methods 348
 - symmetric methods 342
 - symplectic methods 343
 - trees 369
 - variable steps 356
- modified equation
 - parasitic 579
- modified Hamiltonian 343, 375, 589
 - global 344, 353
- modified midpoint rule 171
- modulated Fourier expansion 496
 - exact solution 486, 496
 - Hamiltonian 503
 - multi-frequency 519
 - numerical solution 488, 498

- molecular dynamics 18
- mollified impulse method 476, 554
- momenta 181
 - conjugate 181
 - discrete 206
- moments of inertia 100
- momentum
 - angular 9, 98, 100, 101, 173
 - linear 98, 173
- momentum conservation 600
- Moser–Veselov algorithm 281
- multi-force methods 478
- multi-value methods 609
 - symmetric 611
- multiple time scales 472, 479
- multiple time stepping 316, 475
- multirate methods 316
- multistep methods 567
 - backward error analysis 576
 - G-symplectic 587
 - partitioned 572
 - second order equations 569
 - strictly stable 568, 573
 - symmetric 568, 570
 - symplectic 585
 - variable step sizes 605
- Munthe-Kaas methods 125
- N -body system 13, 98
 - energy-momentum methods 173
- Newton-type iteration 331
- Noether’s theorem 210
- non-resonant frequencies 406
- non-resonant step size 433, 498, 511
- Nyström methods 41, 69, 96, 104
 - symplectic 194
- $O(n)$, orthogonal group 119
- one-leg methods 587
- one-step method 8, 29, 187
 - underlying 573, 609
- optimal control 235
- order 29
 - of a tree 53, 67
 - of symmetric local coordinates 167
 - of symmetric projection 162
- order conditions
 - composition methods 71, 75, 80, 93, 94
 - Crouch-Grossman methods 124
 - Nyström methods 69
 - partitioned RK methods 39, 69
 - processing methods 159
 - RK methods 29, 51, 56, 58
 - splitting methods 80, 92
 - symmetric composition 155
 - symmetrized 177
- ordered subtrees 60
- ordered trees 60
- oriented area 183
- oriented free trees 388
- orthogonal matrices 118
- orthogonality constraints 131
- oscillatory differential equations 21, 471, 531
- oscillatory energy 22, 479, 484, 505, 510, 517, 524
- outer solar system 8, 13, 112
- P-series 68, 214
 - symplectic 217, 219
- parametrization
 - tangent space 117
- partial differential equations
 - linear 262
- partitioned Runge–Kutta methods 38, 102, 148
 - diagonally implicit 149
 - symmetric 148
 - symplectic 193, 208, 231
- partitioned systems 3, 66
- pendulum 4, 5, 110, 181, 185, 188, 367, 396, 593
 - double 233
 - spherical 238, 254
 - stiff spring 526
- perturbation series
 - averaging 459
 - Hamiltonian 406
 - reversible 444
- perturbation theory
 - dissipative 455
 - Hamiltonian 389, 404
 - reversible 437
- phase space 2
- Poincaré cut 16
- Poisson
 - bracket 255, 257
 - flow 261, 265
 - integrators 270, 272, 300
 - maps 268
 - systems 254, 257, 297
- Poisson structures 265
 - canonical 254
 - general 256
- polar decomposition 134
- polynomial invariants 105

- potential energy 181, 237
- precession 12, 26
- processing
 - of composition methods 158
 - order conditions 159
- projection
 - symplectic 259
- projection methods 109, 351
 - standard 110
 - Stiefel manifolds 133
 - symmetric 161
 - symmetric non-reversible 166
- pseudo-inverse of a matrix 116
- pseudo-symplectic methods 436

- QR algorithm 108
- QR decomposition 134
- quadratic invariants 101
 - near conservation 225
- quadratic Lie groups 128
- quantum dynamics 293
- quasi-periodic flow 399
- quaternions 281

- r-RESPA method 318, 475
- Radau methods 34
- rank k matrix manifold 131
- RATTLE 245, 280, 352, 388
- resonance
 - numerical 482, 485, 602
- resonance module 517
- reversibility 239, 311
 - of symmetric local coordinates 168
 - of symmetric projection 163
- reversible maps 143, 144
- reversible methods 343
- reversible perturbation theory 437
 - basic scheme 443
 - Birkhoff normalization 447
 - KAM theory 445
 - perturbation series 444
- reversible systems 143
 - integrable 437
 - perturbed integrable 442
- reversible vector fields 144
- ρ -compatibility condition 145
- ρ -reversible 143
 - maps 144
 - vector field 143
- Riccati equation 134
- rigid body 99, 163, 274, 280, 288, 441, 449
 - Hamiltonian theory 278
- Rodrigues formula 141
- rooted trees 53
- rounding error 322
- Runge–Kutta methods 27, 28, 101, 311, 325, 333
 - ρ -compatibility 145
 - additive 50
 - adjoint method 147
 - implicit 29
 - irreducible 220
 - partitioned 38, 148
 - symmetric 146
 - symplectic 191, 231
- Runge–Lenz–Pauli vector 26

- s-stable 594
- Schrödinger equation 293
 - nonlinear 273
- semiclassical dynamics 293
- separable partitioned systems 231
- SHAKE 245
- simplifying assumptions 96
- sinc function 473, 481
- singular value decomposition 133
- $SL(n)$, special linear group 119, 130
- $\mathfrak{sl}(n)$, special linear Lie algebra 119
- small denominators 406
- $SO(n)$, special orthogonal group 119
- $\mathfrak{so}(n)$, skew-symmetric matrices 119
- spherical pendulum 238, 254
- splitting
 - fast-slow 317
 - Lie–Trotter 47
 - Marchuk 47
 - of ordered tree 370
 - Strang 47, 230
- splitting methods 47, 48, 91, 193, 252, 270, 284, 298, 348
 - ρ -compatibility 145
 - negative steps 82
 - of higher order 82
 - order conditions 80
- $Sp(n)$, symplectic group 119
- $\mathfrak{sp}(n)$, symplectic Lie algebra 119
- stability
 - linear 23
 - long-term 592
- stability function 194
- starting approximations 326
 - order 327
- step size control
 - integrating, reversible 314, 357, 449, 538

- proportional, reversible 310, 313, 356, 449
- standard 303
- structure-preserving 310
- step size function 308, 311
- Stiefel manifold 131
- Störmer–Verlet scheme 7, 9, 39, 48, 189, 270, 318, 349, 386, 472, 586
 - as classical limit 300
 - as composition method 148
 - as Nyström method 41
 - as processing method 159
 - as splitting method 48
 - as variational integrator 208
 - energy conservation 368, 513
 - linear error growth 414
 - symmetry 42, 145
 - symplecticity 48, 190
 - variable step size 308, 309, 312, 313, 315
- Strang splitting 47, 230, 315, 348
- structure constants 286
- submanifold 109
 - symplectic 259
- subtrees
 - ordered 60
- summation
 - compensated 323
- superconvergence 32, 37, 250
- Suzuki’s fractals 45, 46, 153
- switching lemma 76
- symmetric collocation methods 146, 176
- symmetric composition 94
 - of first order methods 150
 - of symmetric methods 150, 154
- symmetric composition methods 149
 - of order 6 156
 - of order 8 157
 - of order 10 158
- symmetric Lie group methods 169
- symmetric methods 3, 42, 143, 144, 342, 612
 - explicit 148
 - symmetric composition 154
- symmetric methods on manifolds 161
- symmetric projection 161
 - existence of numerical solution 162
 - non-reversible 166
- symmetric Runge–Kutta methods 146, 176
- symmetric splitting method 177
- symmetrized order conditions 177
- symmetry 289, 311, 613
 - of Gauss methods 147
 - of Lobatto 147
 - of symmetric local coordinates 168
- symmetry coefficient 57, 67, 72
- symplectic 183, 196, 241
 - B-series 217
 - maps 268
 - P-series 217
 - projection 259
 - submanifold 258, 295
- symplectic Euler method 4, 48, 189, 193, 230, 242, 270, 340, 346, 349, 383
 - as splitting method 48
 - energy conservation 368
 - variable step size 307
- symplectic methods 187, 612
 - as variational integrators 207
 - based on generating functions 203
 - irreducible 222
 - Nyström methods 194
 - partitioned Runge–Kutta methods 193, 208
 - Runge–Kutta methods 191
 - variable step size 306
- symplectic submanifold 259
- symplecticity 244, 585
- Takens chaos 563
- tangent bundle 239
- tangent space 114, 120
 - parametrization 117, 134
- θ -method 147
 - adjoint 148
- three-body problem 321, 390
- time transformation 306, 356
- time-reversible methods 144
- Toda flow 109
- Toda lattice 402, 414, 440, 449
- total differential 186, 196
- total energy 9, 18, 21, 98, 479, 484, 510, 524, 600
- transformations
 - adiabatic 531, 532
 - averaging 458
 - canonical 186
 - reversibility preserving 438
 - symplectic 182, 183, 196, 241
- trapezoidal rule 28, 194, 223, 312
- trees 51, 217, 369
 - bi-coloured 66
 - equivalence class 384
 - ordered 60
 - oriented free 388

- rooted 53
- ∞ -trees 72
- trigonometric methods 473
- triple jump 44, 46, 153
- true anomaly 9
- two-body problem 9, 25
- two-force methods 478
- underlying one-step method 573, 609
- Van der Pol's equation 455
- variational integrators 204
- variational problem 205, 237
- variational splitting 271
- vector fields 2
 - divergence-free 227
 - reversible 143, 144
- Verlet method 7, 39, 48, 189, 270, 318, 472, 513
 - adaptive 309
- Verlet-I method 318, 475
- volume preservation 105, 113, 227, 231
- volume-preserving integrators 228
- weak invariants 109
- work-precision diagrams 150, 153, 156, 157, 334, 336, 482, 604, 605, 608
- W -transformation 235